DOCUMENT RESUME

ED 404 368 TM 026 454

AUTHOR Barron, Sheila I.; Koretz, Daniel M.
TITLE An Evaluation of the Robustness of the NAEP Trend Lines for Racial/Ethnic Subgroups. NAEP TRP Task 3h: Non-Cognitive Variables.
INSTITUTION National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
PUB DATE 20 Dec 94
CONTRACT RS90159001
NOTE 63p.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Statistical Data (110)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Age Differences; *Educational Trends; Elementary Secondary Education; Error of Measurement; *Estimation (Mathematics); Ethnic Groups; *Minority Groups; *Racial Differences; *Robustness (Statistics); Sample Size; *Trend Analysis
IDENTIFIERS *National Assessment of Educational Progress

ABSTRACT

Recent changes in the National Assessment of Educational Progress (NAEP) that lead to its division into a trend assessment and a main assessment jeopardize the information the NAEP can provide about trends, especially the trends for racial and ethnic groups. This study for the Technical Review Panel addressed whether the trend assessment provides overly error-prone estimates for population groups and whether estimates are substantially different than those that would have been obtained had the trend assessment more closely resembled the main assessment. Data from the trend assessment for all its years of administration through 1992 and from the 1984 and 1992 main assessments were used, along with Census data. The combination of smaller samples and the lack of oversampling of minorities results in extremely large confidence intervals for Black and Hispanic means for the trend assessment. To explore systemic differences between the trend and main assessments, differences in the method used to identify minority students, the use of age-defined rather than grade-defined samples, and differences on content and format were studied. Both the (large) differences in ethnic classification and the use of age-defined samples appear to have erratic effects on trend lines, but differences in format and content have little impact. The findings are uncertain primarily because of the large standard errors for the minority results. Recommendations are offered to improve the trend lines. (Contains 7 figures, 9 tables, and 13 references.) (Author/SLD)

C·R·E·S·S·T

National Center for Research on Evaluation, Standards, and Student Testing

Draft Deliverable — December 20, 1994

An Evaluation of the Robustness of the NAEP Trend Lines for Racial/Ethnical Subgroups

NAEP TRP Task 3h: Non-Cognitive Variables

Sheila I. Barron
Daniel M. Koretz

- **UCLA Center for the Study of Evaluation**

in collaboration with:

- **University of Colorado**
- **NORC, University of Chicago**
- **LRDC, University of Pittsburgh**
- **University of California, Santa Barbara**
- **University of Southern California**
- **The RAND Corporation**

National Center for Research on
Evaluation, Standards, and Student Testing
Technical Review Panel for Assessing the Validity of
the National Assessment of Educational Progress

Draft Deliverable — December 20, 1994

An Evaluation of the Robustness of the NAEP Trend Lines for Racial/Ethnical Subgroups

NAEP TRP Task 3h: Non-Cognitive Variables

Sheila I. Barron
Daniel M. Koretz

U.S. Department of Education
National Center for Education Statistics
Grant RS90159001

Center for the Study of Evaluation
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

The work reported here was supported by a grant from the National Center for Education Statistics Contract No. RS90159001 as administered by the U.S. Department of Education and the Lilly Endowment.

The findings and opinions expressed in this report do not reflect the position or policies of the the National Center for Education Statistics, U.S. Department of Education or the Lilly Endowment.

TABLE OF CONTENTS

DRAFT: AN EVALUATION OF THE ROBUSTNESS OF THE NAEP TREND LINES FOR RACIAL/ETHNICAL SUBGROUPS

ABSTRACT

The National Assessment of Educational Progress (NAEP) is only reference available for discussing trends in the achievement of American students where representative samples of students are assessed at relatively frequent intervals. However, relatively recent changes in NAEP that lead to its division into a trend assessment and a main assessment jeopardize the information NAEP can provide about trends, especially trends for "racial/ethnical" population groups. Two questions were addressed in this study: first, whether the trend assessment provides overly error-prone estimates for population groups, and second, whether estimates are substantially different from those that would have been obtained had the trend assessment more closely resembled the main assessment. The combination of smaller samples, and the lack of oversampling of minorities results in extremely large confidence intervals for black and Hispanic means the trend assessment. To explore systemic differences between the trend and main assessments, we investigated differences in the method used to identify minority students, the use of age-defined rather than grade-defined samples, and differences in content and format. Both the (large) differences in ethnic classification and the use of age- rather than grade-defined samples appeared to have erratic effects on trend lines, while differences in format and content appeared to have little impact. These findings, however, are uncertain primarily because of the large standard errors for the minority results. Based on these findings, we offer several recommendations, including oversampling of minorities in the trend assessment and re-evaluating the rigid constancy of the trend assessment.

DRAFT: AN EVALUATION OF THE ROBUSTNESS OF THE NAEP TREND LINES FOR RACIAL/ETHNICAL SUBGROUPS

Sheila I. Barron, The RAND Corporation

Daniel M. Koretz, The RAND Corporation

INTRODUCTION

For more than 20 years, the National Assessment of Educational Progress (NAEP) has been the primary indicator of the academic performance of American youth. It is the only assessment administered frequently to large, nationally representative samples of students in a variety of subject areas.

Although NAEP results are used in many ways, measurement of trends in performance over time has been one of the most important functions the assessment has served. In recent years, measurement of trends for population groups ("racial/ethnic" subgroups) has assumed growing importance.[1] For example, NAEP results presenting differences in the trends among population groups have been instrumental in alerting the public to the gains of black students relative to their non-Hispanic white peers (see Koretz, 1986; Mullis, et. al., 1991) .

For nearly a decade, however, the NAEP's estimates of moderate- and long-term trends have been based on a different assessment than the main

---

1 The terms "race" and "ethnicity" are misleading in this context. Commonly used "racial" and "ethnic" categories are socially conventional classifications that include racial and ethnic components but are not clearly racial or ethnic. For example, ethnically diverse Hispanics are lumped together in a single category, and individuals of mixed white and black ancestry are typically classified as "black" even if their ancestry is as much white as black. Moreover, the "racial/ethnic" classification of individuals is inconsistent from one data source to another. For this reason, we use the neutral term "population group."

NAEP assessment that is used for cross-sectional comparisons, short-term trend estimates, and (with modifications) for the Trial State Assessment. The aspects of the trend assessment that differ from the main assessment are substantial and include smaller samples of students, much sparser item sets, less variation in content and format, and different administrative and reporting procedures.

The differences between the main and trend assessments raise questions about the robustness of NAEP's estimates of trends for population groups. Two distinct questions arise: first, whether the trend assessment is providing estimates for these groups that are overly error-prone, and second, whether estimates from the trend assessment are substantially different from those that would have been obtained had the trend assessment more closely resembled the main assessment. This study considers both by examining the impact of a number of threats to the robustness of the trend estimates for population groups: differences in the sampling of students and items, differences in content and format, the use of age-defined rather than grade-defined samples, and the use of different rules for classifying students into population groups. However, before discussing the details of the study, an overview of NAEP is necessary.

An Introduction To Naep

The separation of NAEP into a main assessment and a trend assessment did not occur until the mid-1980s. The 1986 assessment in reading -- the second using test design and scaling procedures introduced by the Educational Testing Service when it took over operation of NAEP -- produced seemingly anomalous results. Specifically, estimated average reading proficiency dropped sharply at ages 9 and 17. This change, particularly at age 17, was far larger than any of the differences between two assessments since the inception of the reading assessments in 1971 (Beaton and Zwick, 1990). Subsequent analysis suggested that changes in the measurement conditions (i.e., timing and item order) had added an unacceptable amount of error to trend estimates in reading (see Beaton & Zwick, 1990). This lead to the decision to separate NAEP into two assessments (Beaton, 1992): a main assessment, which is intended to document what students can do at a particular time and to monitor short-term trends; and a trend assessment, the primary purpose of

which is to monitor longer-term trends. The main assessment continued to incorporate changes, such as revisions of the population definition, the objectives to be tested, the specific items used. In the trend assessment, on the other hand, every effort has been made to maintain consistency over time. Great care has been taken to maintain the same testing procedures and population definitions in the trend assessment.

Unfortunately the use of the term "trend assessment" has not been entirely consistent. Because the main assessment is used, when possible, to assess short-term trends, it has also been called a "trend assessment." The trend assessment referred to throughout this paper is the assessment from which all results reported in the two *Trends in Academic Progress* (Mullis, et. al., 1991; Mullis, et. al., 1994) reports are based.

The design of the trend assessment differs from that of the main assessment. The main assessment, starting in 1988, has had a focused-balanced incomplete block (focused-BIB) design, whereas the design of the trend assessment is probably best described as a unfocused randomly equivalent groups design. In the main assessment, each student is administered a single test booklet which contains three blocks of items, all in the same subject area. (Restricting blocks administered to a student to a single subject area is what is meant by "focused" BIB; the assessments before 1988 used booklets that included more than one subject area.) The blocks are assigned to booklets so that each block is administered with every other block in at least one booklet. For example, in reading for age 13 in 1988, there were seven blocks, each of which was included in three booklets for a total of seven booklets. These booklets were then spiraled within testing sessions.

The trend assessment began before the main NAEP was changed to a focused design, and the trend assessment has not been changed to a focused design. Thus, examinees in the trend assessment are administered blocks in more than one subject area. In addition, in the trend assessment blocks are not placed in multiple booklets (except for writing and one block in age 9 reading). Table 1 presents an overview of the design of the trend assessment for each age and subject area. (For more information the reader is referred to the NAEP *Technical Reports* (e.g., Johnson and Allen, 1992) which are published after each testing round.)

Insert Table 1 about here.

In the trend assessment, reading and writing are administered to one sample and mathematics and science are administered to another. All students in each sample are administered items from both subject areas. For reading and writing, there are a total of six test booklets at each age. Each booklet contains three blocks of items; either two reading blocks and one writing block or two writing blocks and one reading block. In math and science at ages 9 and 13, three booklets are administered, each of which contains a math block, a science block, and a reading block. (The reading block is not scaled; it is only administered to maintain consistency in administrative procedures across time.) At age 17 there are two booklets; one contains two math blocks and one science block, and the other contains two science blocks and one math block.

The reading /writing trend assessment was first administered in its current form in 1988. It employs a subset of the test booklets that had been used in the 1984 [main] assessment of reading and writing. The subset chosen for the trend assessment includes only a small fraction of the booklets but most of the items from the 1984 assessment. (Because BIB spiraling was used in 1984, items appeared in multiple booklets, and the trend assessment could therefore use most of the items while employing many fewer booklets.) The reading/writing trend assessment has now been administered four times: 1988, 1990, 1992, and 1994. However, the 1994 data had not been released at the time this paper was written.

The math/science trend assessment was first administered in 1986. However, it was not envisioned as a trend assessment at that time. The 1986 assessment, which came to be the trend assessment, was originally a bridge assessment developed to link the pre-ETS math and science assessments to the new math and science assessments first administered in 1986. Analysis showed that the math and science bridge had successfully linked the old and new tests (Beaton, 1986), but in light of the conclusions drawn from the reading anomaly, the decision was made not to use the new math and science assessment to monitor long-term trends. Rather, the decision was made to use the small set of booklets administered as a bridge in 1986 as the trend

assessment in math and science. The math and science trend results using this assessment have been reported for three assessment cycles; 1986, 1990, and 1992. It was also administered in 1994.

The impact of these sampling issues on the robustness of the trend lines is discussed later.

METHODS

Data. Data from three sources were used in this study; the main NAEP, the trend NAEP, and the October Current Population Survey (CPS; Bureau of the Census, Series P-20). Data from the NAEP trend assessment were used for all years in which it was administered through 1992. In order to narrow the study, only reading and math were investigated. Data from the 1984 and 1992 NAEP main assessments were also used. In addition, data from the CPS were employed for the years 1984 through 1993 to obtain estimates of the percent of students at ages 9, 13, and 17 who are below modal grade in school. The CPS and NAEP estimates are not directly comparable because the two databases use different age definitions. However, the CPS data is useful for comparing trends in the percent of students below modal grade for the population as a whole and for population groups.

Scale. NAEP uses a unique scaling method. Proficiency scores for individuals are determined through the use of item response theory (IRT) and multiple imputations, or plausible values, methodology (for more information see Mislevy, Johnson, and Muraki, 1992). The method does not provide a point estimate of each individual's proficiency. Rather, it produces five "plausible values" for each individual drawn from a posterior proficiency distribution that is obtained by conditioning students' responses on a number of cognitive and background variables. This conditioning is designed to offset the effects of measurement error (from the short test length for individual students). This methodology has advantages for estimating aggregate level order statistics and standard deviations. In addition, the standard deviation of a statistic estimated separately using each of the five plausible values provides an estimate of the amount of uncertainty in the statistic due to employing a latent trait scaling methodology.

For some of the analyses conducted in this study the plausible values scale presented problems. In these cases, it was necessary to go to an alternative metric. In cases where some of the examinees were administered a very small number of items, it did not seem wise to rely on unconditioned proficiency estimates or conduct a non-IRT "equating." However, because the same items were administered in each assessment, it was possible to use a probit transformation of the item proportion correct as the metric. The probit of an item's proportion correct is the quantile from the standard normal distribution. For example, an item with a p-value of .5 would have a probit of 0. Using the probits it is possible to compare performance over time on subsets of the items. Also, it is not necessary when using probits to be concerned about changes over time in the NAEP conditioning variables.

RESULTS

Sampling Of Students

The total number of students assessed in the trend assessment, while smaller than the sample of the main NAEP, is sizable. The number of minority students in the trend assessment, however, is relatively small. The smaller number of blacks and Hispanics assessed in the trend assessment, compared to the main NAEP assessment, is due not only to the relatively smaller trend sample sizes, but also to the decision to not over-sample high minority schools in the trend assessment. (Such schools are oversampled in the main assessment.) The combination of these two factors lead to standard errors for minority-group statistics in the trend assessment that are much larger than -- often about double -- the corresponding standard errors in the current main assessment or that were available for trends before the trend assessment was separated from the main assessment.

Given this sampling, only huge changes in the performance of minority groups are statistically significant. Large and educationally important gains may escape detection, and estimates of the magnitude of changes are highly uncertain. This problem is considerable for the trend lines for blacks but is especially severe for Hispanics because of their even smaller sample size.

The amount of change required for significance depends on the population group examined. For the large sample of whites, a difference of

four or more points is generally significant (at a=.05 per comparison[2]). For the black population groups (age 9, 13, and 17) in math, a difference of anywhere from five to more than seven points (depending on the comparison of interest) would be required to show a significant change and, in reading, a difference of anywhere from six point to almost eight points is required. For the Hispanic subgroup, the sample sizes are the smallest and thus the standard errors are the largest. In math, a score difference of anywhere from six to more than eight points is necessary and, in reading, a score difference of anywhere from almost eight points to more than eleven points is required.

The impact of this low power is apparent when the significance and size of trends in group means are compared. During the years that the trend assessment has been separated from the main assessment, only one minority trend line in reading (age 17 for the black subgroup) has shown a significant change from 1992 (using a=.05 per family of comparisons) and that was a 13 point score decrease (see Figure 1). At age 13 in reading, the white subgroup showed a significant increase in mean performance from both 1988 (a five point change) and 1990 (four point change). In math, there were no significant changes from 1992 in minority mean performance. Meanwhile in all ages, there has been a significant improvement in math scores between 1986 and 1992 for the white subgroup. The clearest example of the low power of minority comparisons is at age 17 in math. A score improvement between 1986 and 1992 of four points was significant for the white subgroups whereas a score improvement of nine points for the Hispanic subgroup over the same time period was not.

---

2 The alpha level used in reporting NAEP trend results is typically .05 *per family* of comparisons. Thus, our use of .05 per comparison results in a somewhat smaller difference being required for statistical significance and therefore understates the severity of the problem in reported NAEP results. This was done to avoid the need to explain each family of comparisons.

Insert Figure 1 about here.

One way to get an intuitive feel for the importance of the changes that fail to reach statistical significance (although it can be misleading and should be used carefully) is to recast the NAEP scale in terms of grade equivalents. Using the mean achievement scores as representative of performance in the modal grade and assuming that grade to grade change is constant, a year of instruction corresponds to approximately 11 points on the NAEP proficiency scale. Given this interpretation of the scale, a change in mean performance for a population group of a few points may be meaningful.

Thus, for example, the apparent increase in the mean score of age 17 Hispanics between 1986 and 1992 was 9 points, if true, represents a massive gain -- an improvement of nearly a year's instruction in the space of only 6 years -- but it failed to reach significance. Similarly, the 10-point improvement in reading for age 17 Hispanics between 1980 and 1992 failed to reach significance.

Another informative gauge of NAEP's sampling error is to place a confidence band around trends in the mean scores of minority groups. NAEP is not used simply to determine *whether* minority students have shown improvement; it is used as well to estimate the *amount* of their improvement. One of the more extreme examples is the change in Hispanic reading scores from 1980 to 1990. The observed score increase of 14 points was significant. However, it is important to know not just whether the scores of Hispanics have improved, but also how large any improvement has been, and the NAEP trend data cannot provide an adequate answer to the latter question. The 95% confidence band for the Hispanic score gain extends from 5 points to 23 points.

Sampling Of Items

The size of the standard errors is determined by the amount of sampling error and the amount of measurement error. Thus, inadequate sampling of items could also threaten the robustness of the trend lines. There are two aspects of sampling of items that are important to consider; the overall number of items and the number of items administered to each examinee. The trend assessment involves fewer items overall and drastically fewer items

per examinee than the main NAEP assessment. In Table 2, the 1992 main and trend assessments are used to illustrate the difference in the total number of items administered.

Insert Table 2 about here.

The overall number of items needed to adequately sample the domain depends on the breadth of the domain of interest. If the domains assessed in the trend assessment are more narrowly defined than the respective main assessment domains then the smaller number of items may be reasonable. **(However, for this to be the case, the conclusions based on the trend assessment would have to clearly reflect this difference in domain definitions.)** In addition, smaller item samples may suffice in the trend assessment because that assessment, unlike the main NAEP assessment, is not used to report subscale scores.

The number of items administered per examinee determines, in part, the precision with which the examinees' proficiency is estimated. Fewer items per examinee, all other things being equal, leads to greater measurement error which is reflected in the standard errors. The standard deviation of the plausible values computed by booklet provided an indication of the impact of administering a small number of items to each examinee.

The estimates of measurement error in the trend assessment indicate that, relative to sampling error, measurement error is not a serious threat even when only five items are administered to a proportion of the examinees. Even when the number of items is very small, measurement error in the trend estimates is not very large. The greatest booklet-to-booklet discrepancy in the number of items administered is at age 17 in math, where one-half of the examinees takes 66 items and the other half take five items. In both 1986 and 1990 in math at age 17, the estimate of measurement error in Booklet 84 (I=66) is approximately .2 scale points. The estimate of measurement error in Booklet 85 (I=5) is twice as large but still very small -- approximately .4 scale points.

It is essential to note, however, that NAEP can obtain efficient proficiency estimates for examinees using only a small number of items only because the estimates are obtained through conditioning on background

information about the examinees. Moreover, the smaller the number of items, the greater the reliance on conditioning.

The importance of this reliance on conditioning is illustrated by the results for age 17 math in 1992. In that instance, the relationship between the number of items and the measurement error was reversed: The estimate of measurement error in Booklet 84 (I=66) is approximately .3 scale points and the estimate of measurement error in Booklet 85 (I=5) is approximately .1 scale points. This is a clear example of the drawback of using plausible values methodology. Results are obtained that are clearly contrary to expectation and the methodology is so mathematically complex that it is extremely difficulty if not impossible to determine the cause for the aberrant results.

A second issue concerning sampling of items is related to the inefficient use of student time in the trend assessment. Students are administered three fifteen minute blocks of items as well as a preliminary block of background information. The fifteen blocks of items also begin by asking a number of non-cognitive background questions, so students spend less than the full 15 minutes on cognitive questions. In addition, it is known in advance that some of the administered cognitive items will not be used either because they have been found in the past not to work well, or because they assess math using a calculator, and calculator items are not scaled in the trend assessment. Furthermore, there are several blocks of items which are not scaled at all.

This problem is most severe in the math trend assessment. Table 3 presents the number of math items in each booklet that were scaled in 1992. At ages 9 and age 13, one third of the testing time of all students is wasted by administering reading blocks which are not scaled in order to maintain consistency in administration. Thus, maintaining any context effects due, for example, to taking a math block after a reading block. In addition, for one third of these students, a second block of items is almost entirely wasted by administering calculator items which are not scaled. At age 17, where one of the two booklets contains a single math block made up almost entirely of calculator items, approximately one half of the students assessed have scaled scores based on only five math items.

Insert Table 3 about here.

Population Groups

Population group membership is determined differently in the trend assessment than it is in the main NAEP assessment. In the main assessment, examinees are placed into population groups using self-reported information whenever possible (when that information is available and usable) and using the exercise administrator's observation only in the small number of cases where self-reports cannot be used. In contrast, in the trend assessment, only the exercise administrator's observation is used to identify population groups. We examined the consistency of classification between methods and the impact on the trend lines of means of classification. Because of differences in the accuracy of self-reported information for students of different ages, special attention was given to differences in the results for the three age levels assessed by NAEP.

Consistency of classification. The first step in assessing the importance of the method used for determining population group membership was to crosstabulate the two population-group variables. The variable used in the trend assessment, called "observed race" in much of the NAEP documentation, is simply the exercise administrators judgment as to the racial/ethnical background of the each student. The variable used in the main assessment, called "derived race" because it combines information from multiple sources, gives priority to student reported information and only uses the exercise administrators judgment if the student omits the race/ethnicity information or answers a relevant question with multiple responses. Both variables use mutually exclusive categories labeled black, white, and Hispanic[3]. However, because Hispanic students may belong to any racial group, it is necessary to decide which variable takes precedence in the case of Hispanics. The decision rule in both the trend and main assessment is that students who are ethnically Hispanic should be classified as Hispanic regardless of their race. That is,

---

3 There are other population group categories as well (i.e., Asian, American Indian). However, due to the small sample sizes, they are not used in NAEP reporting of long-term trends.

students who are classified as Hispanic are counted as neither white nor black.

The two classification systems are highly consistent for blacks and whites but strikingly inconsistent for Hispanics. The variable used in the main assessment--"derived race"--classifies far more students as Hispanic than does the "observed race" variable used in the trend assessment. Although some of the students classified differently by the two variables are black (accorded to the observed race variable), the main source of inconsistency is examinees who report that they are Hispanic but are considered white (not Hispanic) by the exercise administrators, that is, by observed race. The most extreme instance is at age 9, where only 40 percent of the students classified as Hispanic by derived race are also classified as Hispanic by observed race (Table 4). The percent agreement increases with age but remains a problem at all ages: it rises to 62 percent at age 13 (Table 5) and 69 percent at age 17 (Table 6).[4]

Insert Table 4-6 about here.

The decrease in disagreement between the two variables as age increases could indicate that younger students more often misclassify themselves because of not understanding one or both of the questions. To explore this possibility, we examined the consistency of responses to two background questions, one of which asked about race (and included the option of "white (not Hispanic)" and the second of which asked students which Hispanic group (Mexican, Mexican American, Chicano, Puerto Rican, Cuban, or other Spanish/Hispanic) they belong to. We looked at the responses to these questions for students who identified themselves as Hispanic but were identified as white by the test administrator. Because the wording of these questions in the trend assessment was not clear-cut, this analysis was conducted using the main assessment data.

Younger students are indeed more likely to answer these two background questions inconsistently. The percent of these students who

---

4 Crosstabs were computed for the reading samples in 1988, 1990, and 1992 and for the math samples in 1986, 1990, and 1992. However, because there was not a identifiable consistent difference in identification across years or subject areas only the 1992 math results are presented.

responded that they are white (Not Hispanic) in response to the race background questions and that they are Hispanic (Mexican, Mexican American, Chicano, Puerto Rican, Cuban, or other Spanish/Hispanic) in response to the ethnicity background question decreases as age increases. For example, in reading in 1992, the about 36 percent of age 9 examinees with observed race equal to white and derived race equal to Hispanic who answer 'white (Not Hispanic)' to the first background question but choose an Hispanic option in the second background question. The corresponding percentages were 24 and 13 at ages 13 and 17, respectively.

Similarly, the percent of misidentified students (with derived race=Hispanic and observed race=white) who consistently identify themselves as Hispanic increases with age. At age 9, 38% (46% if other is considered a consistent option) of the misidentified students identified themselves as Hispanic in response to both questions. At age 13, 56% (66%) of the misidentified students consistently identified themselves as Hispanic and, at age 17, the percent was 74 (83%).

Although the classification of Hispanics in the NAEP trend data is seriously inconsistent with that in the main NAEP assessment, this analysis does not clearly suggest that either method is sufficient, especially for younger examinees. On the one hand, the arguments against reliance on judgments by the test administrators are clear: they will typically have only limited and potentially misleading information, such as appearance and surname. On the other hand, the inconsistencies in self-reports shown here suggest that self-reports are also suspect, at least for students at age 9. Further research is needed to explore the validity of alternative classification methods.

The impact of classification inconsistency on trend estimates. Given the sizable discrepancies between the classification systems used in the trend and main NAEP assessments, it is important to investigate the practical impact this has on the observed trend lines.

In some cases, the classification does have appreciable effects on the means for blacks and Hispanics, and it appears that they may affect trends (Figure 2). However, the effects of the difference in classification are both

erratic and small relative to the standard error of the group means (which are large because of the small minority samples in the trend assessment)[5].

Insert Figure 2 about here.

In contrast, the effects of the different classifications for non-Hispanic whites are consistent although very small. Using derived race (the variable used in the main assessment) results in white means that are approximately one point higher than the means using the observed race variable. This is the effect that one would expect if a proportion of the Hispanic group, which on average scores lower than whites, is included in the white subgroup because of misclassifications by test administrators. And, even though the proportion of the Hispanic group being included in the calculation is substantial (i.e., 40% of examinees self-identified as Hispanic), the impact on the mean for the white subgroup is small because of the much larger number of white students in the sample.

Content

An investigation of the impact of the relative weight given to different types of content in the trend assessment was undertaken in reading and math. The trend assessments are based on content frameworks that were developed for either the 1983-84 assessment (reading) or the 1985-86 assessment (math).

---

[5] Apart from large standard errors, there is a technical reason to be cautious in the interpretation of these plots. The conditioning model used to generate the plausible values has not always included the derived race variable. In years where derived race is not included, the estimates of the means calculated by derived race are statistically biased estimates of the population values. The amount of bias is related to the covariance between derived race and the variables that are included in the conditioning model. Rather than attempt to estimate the size of the bias, we replotted the trends using probit-transformed percents correct, which are not dependent on the conditioning model. The trends broken down by both observed race and derived race in the probit metric support the conclusions reached using the plausible value metric.

Substantial changes have occurred since the development of these frameworks in the objectives in which content experts believe teachers should be striving to teach. NAEP will continually have to struggle with the conflict between assessing the objectives currently considered important and maintaining consistency over time in what is assessed. For example, should one conclude that achievement in a subject area has actually gone down when the indicator of this trend is performance on items developed to test objectives that are no longer considered of primary importance by educators?

In recent years, NAEP has take fundamentally different approaches to this tension in designing the trend and main assessment. The current practice is to make the changes called for by content experts and supported by the National Assessment Governing Board to the frameworks used in the main assessment but leave the trend assessment frameworks undisturbed. This ensures that a common score scale over time is maintained in the trend assessment. However, the practical result of this practice is that the content frameworks used in the trend assessment are quite different than those employed in the main assessment.

It appears clear from our dealings with consumers of NAEP, even well-informed consumers, that many people assume that the frameworks published periodically by NAEP are the frameworks used in the trend assessment. In fact, this misunderstanding is supported by the NAEP documentation itself, however inadvertently. On page one of the 1990 Science Objectives documentation (March 1989) is the following sentence: 'Previous assessments in science were conducted during the school years ending in 1970, 1973, 1977, 1982, and 1986; thus the 1990 assessment of students at grades 4, 8, and 12 and at ages 9, 13, and 17 will provide a view of science achievement that spans 20 years.' One page 5 of the *Overview of NAEP Assessment Frameworks* (March 1994) are the following two paragraphs under the heading Trend Assessment:

> Parallel tracks of assessment are run to maintain the stability required for measuring trends while still introducing innovations. Approximately half the NAEP items used in each subject area are reused in later assessments to measure change over time. To keep pace with developments in assessment methodology and research about learning in each subject area, NAEP updates the other half with each

> successive administration and releases the items not designed for reassessment for public use.
>
> Trend items are selected based on their representativeness in view of the framework objectives and on psychometric characteristics obtained from the assessment to ensure the released and unreleased parts of the assessment are as equivalent as possible in difficulty and other measurement considerations.

These paragraphs apparently refer to the practice in NAEP of attempting to maintain short-term trend lines using the main assessment data. However, the reader is left unaware that the trend assessment referred to in this document is completely separate from the trend assessment used to obtain data for the *NAEP Trends in Academic Progress Report* (1991) and the *NAEP 1992 Trends in Academic Progress Report* (1994).

We do not advocate the use of rapidly changing content frameworks in the assessment of achievement trends over time, but it is important to call attention to the differences in the frameworks used in the trend and main assessments and to investigate the impact of these differences on the reported trend lines. The latter is explored in the subject areas of math and reading.

Mathematics. The content of the trend assessments in math differs from that of the main assessment in a number of ways. One fundamental differences is a shift away from numbers and operations in the main assessment. In the trend assessment, roughly 50% of the items scaled at each age are numbers and operations items. In the main assessment in 1992, the percent of the items that are classified as numbers and operations in 1992 was 40, 32, and 24 in grades 4, 8, and 12, respectively (see Table 7).

Insert Table 7 about here.

Although it is not possible to estimate what the trends would have been for content that was not assessed, it is possible to examine the variability in the trends computed separately for different content categories. Using data from the trend assessment, separate trend lines (in the form of the average of the probit transformations of the p-values) were plotted for each content classification. Some content categories were represented by very few items in the trend assessment, and results are not plotted for any content area with less than five scaled items. In addition, a few items that were scaled in one or

more years but not in all years were not used in the computation of the trend lines. In other words, only items scaled in all three assessment years were included in the computation of the probit trend lines.

In the math trend assessment, there is very little evidence of differential trends by content area for any of the population groups for which trends are reported[6]. The trend lines for each content area were plotted for all examinees and separately for each of the three largest population groups. Because the relationship between content area trends does not differ substantially by population group, only the overall trends are presented (see Figure 3). At Age 9, the only content area which showed a trend different than the overall trend was Data Organization/Interpretation. This difference was consistent across population groups, so it would have no appreciable impact on relative trends by population group. For all three population groups at Ages 13 and 17, all of the content areas (with I >= 5) had trends reasonably consistent to the overall trend.

Insert Figure 3 about here.

Reading. The content in reading is broken down by reading objective in the trend assessments and purpose of reading in the main assessment. There are three categories of reading objectives in the trend assessment: 1) reading to derive information, 2) integration and application, and 3) evaluation and reaction. Items that do not fit well into any of these three categories are placed in a fourth miscellaneous category. There are also a few items that are not classified. The main assessment in reading divides items according to reading purpose;: 1) reading for literary experience, 2) reading to gain information, and 3) reading to perform a task. There appears to be rough mapping of the trend objectives to the main assessment purposes. However, the shifts in the main assessment have not been merely assigning different relative weights to the various content areas but rather a fundamental shift in how content in reading is delineated. Thus examining how trends vary according to the objectives in the trend assessment is a very conservative estimate of how the trends might vary if the items looked more like the main assessment items.

---

[6]Population group membership was identified using the observed race variable.

Using data from the trend assessment, separate trend lines (in the form of the average of the probit transformations of the p-values) were plotted for subsets of the items determined by the reading objective. The trend lines were plotted for all examinees and separately for whites, blacks, and Hispanics. Once again, only items scaled in all assessment years were included in the computation of the probit trend lines.

There is a tendency for items classified as "evaluate and react" to show a greater positive change over time than is shown by items assessing the other two objectives. This is true especially at age 17. However, because the evaluate and react objective is represented by so few items in the trend assessment, it is not possible to determine whether this is a trend that is unique to these items or one that would generalize to other items designed to assess this objective. Figure 4 presents the overall content area specific trends; within-group trends are not shown because the relationship between the content area trends does not differ substantially by population group.

Insert Figure 4 about here.

Item Format

In addition to fundamental shifts in the content specifications for the main and trend assessments, there has been a shift in the main assessment toward greater use of item formats other than multiple choice. This shift in the main assessment reflects a shift in public attitude towards assessment. To the extent that different item formats tap different aspects of proficiency, however, the NAEP trend lines may not be robust against changes in item format.

The trend assessment in reading is made up almost entirely of multiple choice items whereas the current main assessment in reading is comprised of approximately one-half multiple choice items and one-half constructed response items. The trend assessment in math contains a fair number of non-multiple choice items that are probably best called short open-ended. but does not include more extensive constructed response items. In the main assessment, on the other hand, a large proportion of the items are short constructed response items, and there are also a number of extended constructed response items.

Mathematics. There is some evidence suggesting differential trends for short open-ended items and multiple choice items. Performance over the time period examined here tends to be relatively constant for the short open-ended items whereas it has been increasing on the multiple choice items (see Figure 5). Although not presented here, this difference in trends between formats is replicated for each of the three main population groups. Given that the same trend appears at all ages, sampling error is probably not a serious threat to this conclusion.

However, given the relatively small number of items, the generalizability of this finding to other items of these types is uncertain. It is important to note that the open-ended items in the trend assessment (often fill in the blank) are typically quite short -- they are dichotomized for scaling purposes -- thus the trend in performance of examinees on the open-ended items in the trend assessment may not be a reasonable estimate for how the trend on more extensive items might appear. These results suggest that if more weight had been given to open-ended items of the form included in the trend assessment, the significant improvement in math achievement evidenced between 1986 and 1992 would not have appeared or would have been smaller. The impact of including more substantial open-ended items remains unclear.

Insert Figure 5 about here.

Reading. NAEP classifies the trend items as either multiple choice or open-ended. However, the open-ended items in the trend assessment are of two distinct types: 1) open-ended items that require performing a task (which will be called non-constructed response (non-CR) open-ended items), and 2) items that require writing out an answer (CR items). The vast majority of the reading trend assessment items are multiple choice. Due to scoring inconsistencies, the CR items in the trend assessment were not included in the final scaling in 1988. Because the analysis reported above for the reading objectives only included items scaled in all years, there were no constructed response items included in that analysis. However, if 1988 is excluded, it is possible to include constructed-response items in an analysis based on a transformation of the p-values for common items.

Overall and for each of the three population groups, there has been a greater increase in the scores on constructed response items between 1984 and 1992 than there was for scores overall (the average probits calculated by item type for each age level are presented in Figure 6). This is most clear-cut at age 17, where there are more CR items and the increase has been steady across assessment cycles. Thus, the finding that there has not been a significant change in average reading achievement for students ages 9, 13 and 17 for the time period 1984 to 1992 may have been different if the relative weight given to constructed response items had been greater.

Insert Figure 6 about here.

In conclusion, achievement trends appeared to be much more sensitive to item format rather than to content classification/reading objective. Given the small number of non-multiple choice items included in the trend assessment, it is quite conceivable that the observed trends would be different if more weight were given to open-ended items.

Age-Defined vs. Grade-Defined Populations

Historically, NAEP has reported results for populations defined in terms of their age. Three populations were chosen: age 9, age 13, and age 17. After ETS became the NAEP contractor, the reporting focus for the main NAEP results changed to grade-defined populations. Thus, one way in which the populations tested in the two assessments differ is that the main NAEP results are most often reported for populations defined in terms of grade, whereas the trend results (except writing) concern populations defined in terms of age. Trends in achievement over time may differ for these two partially overlapping populations. More important for present purposes, the relative trends shown by population groups may differ between age- and grade-defined samples.

There has been a gradual change over time in the average age of students in a particular grade (and therefore in the grade distribution of students of a particular age). Table 8 presents estimates of the percent of students below modal grade for each year the trend assessment was given. These changes are due to changes in the date by which students must turn five in order to enter Kindergarten in a given year, changes in the voluntary holding out by parents of children old enough to enter school, and changes in

the in-grade retention practices of schools. As a result of these changes, the grade-defined samples assessed by the main NAEP have become older across recent assessments, and the age-defined samples tested by the trend assessment have included an increasing percentage of students below the modal grade for their age.

Insert Table 8 about here.

Changes in the grade distribution of same-age students varied across population groups (e.g., whites and blacks). Table 9 presents estimates of the percent of students below modal grade separately for non-Hispanic whites, blacks, and Hispanics. The change in percent of students below modal grade has been most pronounced for white students, although in absolute terms, the white subgroup still has a lower proportion of students below modal grade.

The impact of changes in the grade distribution of same-age students is not obvious. There is some evidence from research on voluntary holding-out and in-grade retention that older students come to resemble the other students in their grade rather than gaining an advantage because they are older (Shepard & Smith, 1989). Thus, all other things being equal, one might expect increases in the percent of students below modal grade to decrease overall achievement for populations defined in terms of age. However, overall changes in the age-composition of students at a particular grade may be a result of -- and may contribute to -- pushing down of the academic curriculum to lower and lower grades. Thus the impact of these two influences combined could result in increases, decreases or no change at all in the achievement of *age-defined populations*. They would, however, both be expected to contribute in increases in average achievement for *grade-defined populations*. That is, if the population of students at a given grade is both older and has a curriculum that is more advanced (e.g., the fourth grade curriculum that looks more like the fifth grade curriculum of the past) the average achievement of students sampled from that population is likely to be higher than the average achievement to students in that grade in the past.[7]

---

[7] Depending, of course, on the sensitivity of the assessment instrument to these changes.

Insert Table 9 about here.

Changes in the mean achievement gap between majority and minority students may, at least in part, be related to differential changes in the grade distribution of the students. This is especially an issue to the degree that some students do not score well because they have not been presented with some portion of the material on the assessment. In other words because they are below modal grade they have not had an opportunity to learn some portion of the material as it is not presented until the modal grade.

Data from the trend assessment in reading were used to examine the relationship between trend lines for age-defined populations and grade-defined populations. In the reading/writing sample, data is collected on both age-eligible and grade-eligible samples[8], and it is possible to compare the trend lines across the two samples. Figure 7 presents the trends for both age- and grade-defined samples for the three population groups.

Insert Figure 7 about here.

As expected, the trends for the grade-eligible samples are consistently higher than the trend for the age-eligible samples. However, there is not a clear tendency for the gap between the two trend lines to get larger over time as would be expected as more age-eligible students are below modal grade in later years.

SUMMARY OF RESULTS

This paper started out by posing two questions:

1. Is the trend assessment providing estimates for population groups that are overly error-prone? For blacks and Hispanics, the answer to this question is a definite yes. The combination of smaller total samples, compared to the main assessment, and the lack of oversampling of minorities results in

[8] The reading/writing trend assessment is administered to both age-eligible students and grade-eligible students because the reading trend is reported for age-defined populations and the writing trend is reported for grade-defined populations.

confidence intervals for minority means in the trend assessment that are extremely large.

2. Are estimates from the trend assessment substantially different from those that would have been obtained had the trend assessment more closely resembled the main assessment? Unfortunately, the fact that the answer to the first question is yes, makes answering the second question difficult. That is, the large standard errors of results pertaining to minority students clouds the answer to this question. However, it is possible to provide tentative answers to this question.

First, the findings suggest that format differences did affect overall trends but probably did not much influence relative trends among population groups. However, it is important to note that the small sample sizes combined with the small numbers of non-multiple choice items made firm conclusions about differential trend for population groups impossible. Overall, there was some evidence suggesting that the trend lines would be different if the diversity in item format had been greater (more like the main assessment) in both subject areas we investigated. In math, open ended items showed less improvement over time than the multiple choice items. In reading, constructed response items showed greater improvement over time than the multiple choice items or open-ended items.

Second, the means used to identify population groups caused large differences in the identification of Hispanics and created differences, albeit erratic, in the minority trend lines. The disagreement rate between the classification methods drops as age increases, apparently because of a decrease in the error of self-reports. However, even at age 17 the disagreement in who is classified as Hispanic is substantial. Thus, much of the disagreement between classification methods at age 9 and age 13 and almost all of the disagreement at age 17 appears to be due to exercise administrators misidentifying Hispanic examinees as white.

On the other hand, using the trend assessment data, there is little evidence that differences in content (content classifications in math or the reading purpose classifications in reading) had much affect on the trends. In math, the trends plotted by content category for population subgroups mirrored quite closely the respective overall trends. In reading, there was some

evidence to suggest that population groups showed more improvement on reaction and evaluation reading items than was evidenced in other areas of reading. However, given the very small number of items of this type in the trend assessment, this finding may be specific to the few items present.

Finally, the use of populations defined in terms of age rather than grade in the trend assessment has an impact on the location of the trend line but does not appear to have a consistent impact on the size of the majority-minority achievement gap.

CONCLUSIONS

NAEP is only reference available for discussing trends in the achievement of American school children that is based on representative samples and assesses students at relatively frequent intervals. However, relatively recent changes in NAEP that lead to its division into a trend assessment and a main assessment may seriously be jeopardizing the information NAEP can provide about trends, especially trends for "racial/ethnic" subgroups. In addition, the weaknesses in the trend assessment are not widely known because the design and methodology used is often confused with that of the main assessment.

NAEP is currently working on a new trend assessment to replace the present one (see Zwick, 1992). Some of the problems with the current trend assessment noted here will most likely be eliminated when the new trend assessment is in place. For instance, the almost exclusive use of multiple choice items will most likely not be continued in the new trend assessment. However, a new trend assessment will not solve several of the fundamental problems brought up in this study. For example, reliable estimates of trends for minorities will require a substantial change in sampling, one which might require reallocating resources from the main to the trend assessment. Moreover, interpretation of the trend assessment will remain problematic as long as the differences between the main and trend assessments are not made clear to NAEP's audiences.

Based on the findings of this study we have several recommendations. First, sampling in the trend assessment should reflect open discussion about the acceptable size of standard errors for minority group means. Over-

sampling of high-minority schools (as in the main NAEP) or, preferentially, of minority students within schools should be conducted in order to obtain a clearer, more reliable, picture of the trends in achievement for minority students. Both the lack of research on the impact of a heavy reliance on conditioning and plausible values and the inverted relationship noted above between number of items and (conditioned) standard errors suggest that it is unwise at present to continue relying on this method as a surrogate for sufficient minority-group samples.

Second, the ultra-conservative approach to assessing trends that resulted from 1986 reading anomaly should be re-evaluated. ETS concluded based on a study of the reading anomaly that 'When measuring change, don't change the measure.' However, an alternative interpretation of the reading anomaly is that it occurred because changes were made in the measurement instruments without adequate checks built in for making scaling adjustments. Thus, and alternative lesson is this: When changing the measure embed in the design multiple means of checking that the scale has been preserved. The decision to never change the measurement instruments in the trend assessment has led to the various difficulties noted above, and the alternative approach of allowing modest change but building in mechanisms to preserve the scale might avoid or lessen them. For example, ETS's approach led to gross inefficiencies in the use of student time. Specifically, the continued use of reading blocks in the trend assessment of math and science, the continued administration of math calculator items that are not scaled, and the continued administration of items in all content areas that have been found in the past to not be good items and thus are not scaled. A one-time bridge study could replace these blocks and items with items and blocks that are known to work well.

Third, the division between the trend assessment and the main NAEP assessment should be made clearer. It is true that most of the differences between the main assessment and the trend assessment are documented, but it remains unclear--and unrecognized by many users of NAEP results. Given the complexity of NAEP documentation, the multiple uses of the term "trend assessment," and the use of similar scales in the main and trend assessments, it is not at all surprising that most people are unaware of the differences. One suggestion that has been made for solving this problem is to

rename the NAEP trend assessment to something that makes very clear that it is a separate assessment from the main NAEP assessment[9] (e.g., National Assessment of Long-Term Trends).

The fourth and final recommendation follows directly from the third. An open discussion of the long range plan for assessing achievement trends should be held. A consensus should be built on the circumstances under which changes in the trend assessment should be made and the best methodology for maintaining a score scale across time without losing the efficiency needed to maximize the reliability of the trend estimates. We believe that if it were widely understood in the measurement and education communities that the trend assessment does not use the frameworks used in the main assessment and does not balance the use of multiple item formats in the way that the main assessment does, that there would be a public demand for a strategy for assessing trends.

We feel strongly that the National Assessment of Educational Progress, which implies by its very name the assessment of trends, ought to stand as a model for assessing educational trends. Assessing change across time is one of the most difficult tasks in measurement, and NAEP ought to be shining a bright light on both the difficulties involved and the promising avenues for surmounting these difficulties. It was a disappointment to find that the trend assessment is, in many ways, the poor cousin to the main NAEP assessment. And, rather than shining a light on the difficulties inherent in the task of measuring trends over time, the issues are effectively buried.

[9] This suggestion was made by Eva Baker.

References

Beaton, A. E. (1986). *The NAEP 1983-84 Technical Report* (No. 15-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Beaton, A. E. & Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics*, 17, 95-109.

Beaton, A. E. & Zwick, R. (1990). *The effect of changes in the national assessment: Disentangling the NAEP 1985-86 reading anomaly* (Report No. 17-TR-21). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

U.S. Bureau of the Census. (various years). Current Population Reports, Series P-20. *School Enrollment- Social and Economic Characteristics of Students: October 1983 (through 1992)*, U.S. Government Printing Office, Washington DC.

Educational Testing Service. (March, 1989). *Science Objectives: 1990 Assessment* (No. 21-S-10). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Johnson, E. G. & Allen, N. L. (1992). *The NAEP 1990 Technical Report* (No. 21-TR-20). Educational Testing Service, National Assessment of Educational Progress, Princeton, NJ.

Koretz, D. (1986). *Trends in Educational Achievement.* Congressional Budget Office.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP.*Journal of Education Statistics, 17,* 131-154.

Mullis, I. V. S., Dossey, J. A., Campbell, J. R., Gentile, C. A., O'Sullivan, C., & Latham, A. S. (1994). *NAEP 1992 trends in academic progress: Achievement of U.S. students in science, 1969 to 1992, mathematics 1973 to 1992, reading 1971 to 1990 and writing, 1984 to 1990.* Washington, DC: National Center for Educational Statistics.

Mullis, I. V. S., Dossey, J. A., Foertsch, M., Jones, L., & Gentile, C. (1991). *Trends in academic progress: Achievement of U.S. students in science, 1969-70 to 1990; mathematics 1973 to 1990; reading 1971 to 1990; and writing, 1984 to 1990.* Washington, DC: National Center for Educational Statistics.

Shepard, L. A. & Smith, M. L. (1986). Synthesis of research on school readiness and kindergarten retention. *Educational Leadership, 44,* 78-86

White, S. (March, 1994). *Overview of NAEP Assessment Frameworks* (NCES 94-412). Washington, DC: National Center for Education Statistics, U.S. Department of Education, Office of Educational Research and Improvement.

Zwick, R. (1992). Statistical and psychometric issues in the measurement of educational achievement trends: Examples from the National Assessment of Educational Progress. *Journal of Education Statistics, 17,* 205-218.

APPENDICES

Table 1
Design of the NAEP Trend Assessment

Reading and Writing Trend Samples (print administration)

Age 9/Grade 4
There are three writing blocks (one containing a single prompt and two containing two prompts). There are nine reading blocks. Also there is one block which involves a combination of reading and writing items. Six booklets are formed each of which contain three blocks of items with at least one reading block and at least one writing block. Only one reading block in presented in more than one booklet (Block BR is presented in two booklets). Thus, in reading there is very little overlap of items across booklets.

Age 13/Grade 8
There are four writing blocks (two of which contain one prompt and two which contains two prompts). There are ten reading blocks. Six booklets are formed each of which contain three blocks of items with at least one reading block and at least one writing block. In reading, there is no overlap of blocks (or items) across booklets.

Age 17/Grade 11
There are four writing blocks (two of which contain one prompt and two which contains two prompts). There are ten reading blocks. Six booklets are formed each of which contain three blocks of items with at least one reading block and at least one writing block. In reading, there is no overlap of blocks (or items) across booklets.

Although the reading/writing trend samples are age/grade samples, only age eligible students are used in the reading trend and only grade eligible students are used in the writing trend.

Science and Math Trend Samples (paced audiotape administration)

Age 9
There are three science blocks and three math blocks. Three booklets are formed each containing 1 science block, 1 math block, and 1 reading block. The reading block is not used (it is only administered to maintain consistency in administration procedures across time).

Age 13
There are three science blocks and three math blocks. Three booklets are formed each containing 1 science block, 1 math block, and 1 reading block. The reading block is not used (it is only administered to maintain consistency in administration procedures across time).

Age 17
There are three science blocks and three math blocks. Two booklets are formed one containing 2 science blocks and 1 math block, and the other containing 1 science block and 2 math blocks..

Because the administration is paced with an audiotape, all examinees in a session are given the same test booklet. Thus, spiraling is done at the level of session.

Table 2
Number of Scaled Items in the 1992 NAEP Trend and Main Assessments

Mathematics

	Trend	Main
Age 9/Grade 4	55	155
Age 13/Grade 8	80	183
Age 17/Grade 12*	71	179

Reading

	Trend	Main
Age 9/Grade 4	102	85
Age 13/Grade 8	103	134
Age 17/Grade 12*	94	144

*Age definitions and modal grades differ in the two assessments. The trend assessment uses an age definition based on the school year and the modal grade is 11. The main assessment uses an age definition based on the calendar year and the modal grade is 12.

Table 3
Number of Scaled Items per Booklet
1992 NAEP Trend Assessments in Math

	Booklet		
	91	92	93
Age 9	24	5	26
Age 13	36	8	36
	84	85	
Age 17	66	5	

Table 4
Row Percents
1992 Age 9 Mathematics Trend Assessment

Derived Race	Observed Race: White	Black	Hispanic	Other	N
White	97	0	2	1	4829
Black	4	94	1	1	966
Hispanic	40	16	40	4	1221
Other	38	10	5	47	309

Table 5
Row Percents
1992 Age 13 Mathematics Trend Assessment

Derived Race	Observed Race: White	Black	Hispanic	Other	N
White	99	0	0	1	4149
Black	1	97	1	0	810
Hispanic	24	12	62	2	645
Other	22	3	8	68	305

Table 6
Row Percents
1992 Age 17 Mathematics Trend Assessment

Derived Race	Observed Race: White	Black	Hispanic	Other	N
White	100	0	0	0	3295
Black	2	97	0	0	498
Hispanic	22	6	69	3	366
Other	17	3	5	76	200

Table 7
Percent of Items in Each Content Catergory
1992 Main and Trend Math Assessments

Main Assessment

	Grade 4	Grade 8	Grade 12
Numbers & Operations	40%	32%	24%
Measurement	20%	17%	16%
Geometry & Spatial Sense	17%	20%	18%
Data Analysis, Statistics, & Probability	12%	15%	16%
Algebra & Functions	11%	16%	26%

Trend Assessment

	Age 9	Age 13	Age 17
Numbers & Operations	45%	53%	48%
Measurement	23%	17%	10%
Geometry	2%	9%	27%
Data Org./Interpretation	21%	13%	8%
Relations/Functions	4%	4%	11%
Fund. Methods	6%	4%	7%

Table 8

Percent of students Below Modal Grade

	Age=9	Age=13	Age=17
1984	23	28	27
1986	26	29	28
1988	27	30	29
1990	27	31	33
1992	27	30	34
1992-1984	4	3	7

Source: Current Population Survey

Note: Numbers presented are three year rolling averages.

Table 9

Percent of students Below Modal Grade by Population Group

Age 9

	White*	Black	Hispanic
1984	21	30	28
1986	24	34	31
1988	26	33	31
1990	25	32	32
1992	26	33	27
1992-1984	5	3	-1

Age 13

	White*	Black	Hispanic
1984	23	40	44
1986	24	40	43
1988	27	40	41
1990	27	45	40
1992	28	37	38
1992-1984	5	-2	-5

Age 17

	White*	Black	Hispanic
1984	22	41	48
1986	23	42	42
1988	24	42	45
1990	27	49	52
1992	27	52	53
1992-1984	5	11	5

*Not Hispanic

Source: Current Population Survey

Note: Numbers presented are three year rolling averages.

Figure 1

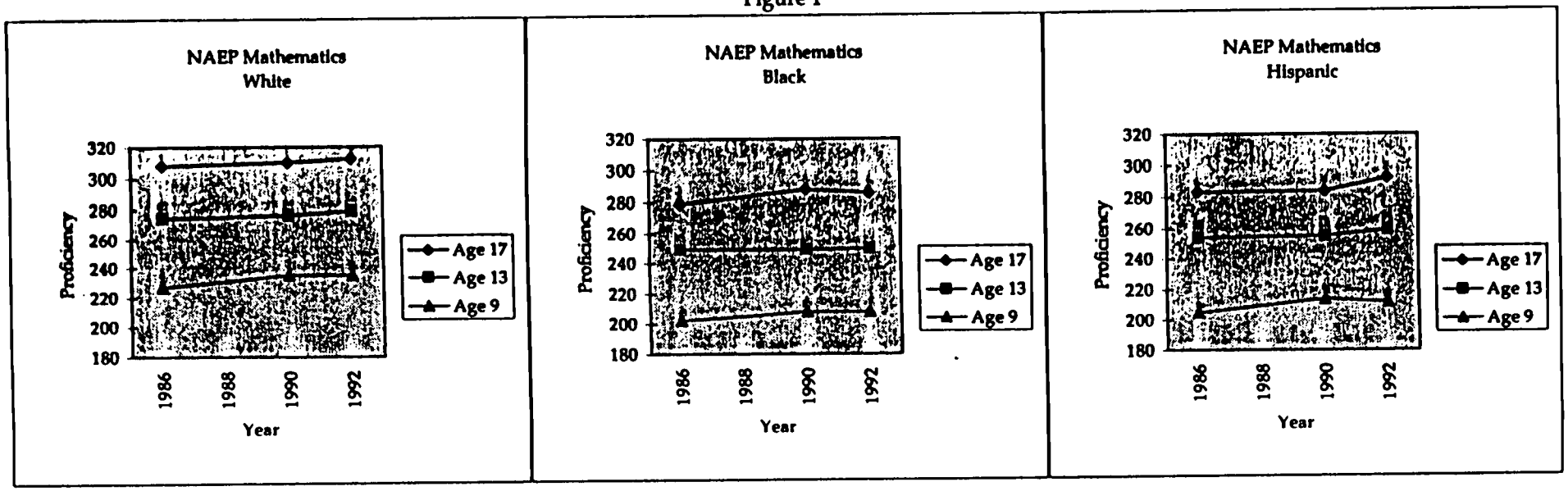


Figure 1 (cont.)

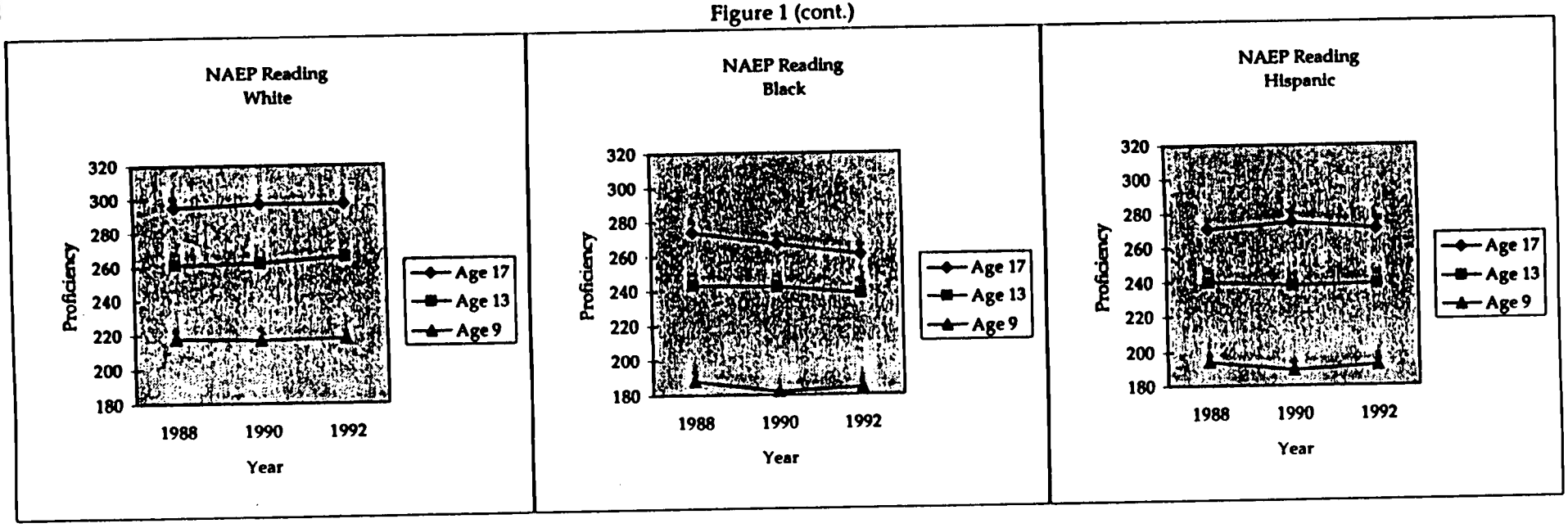
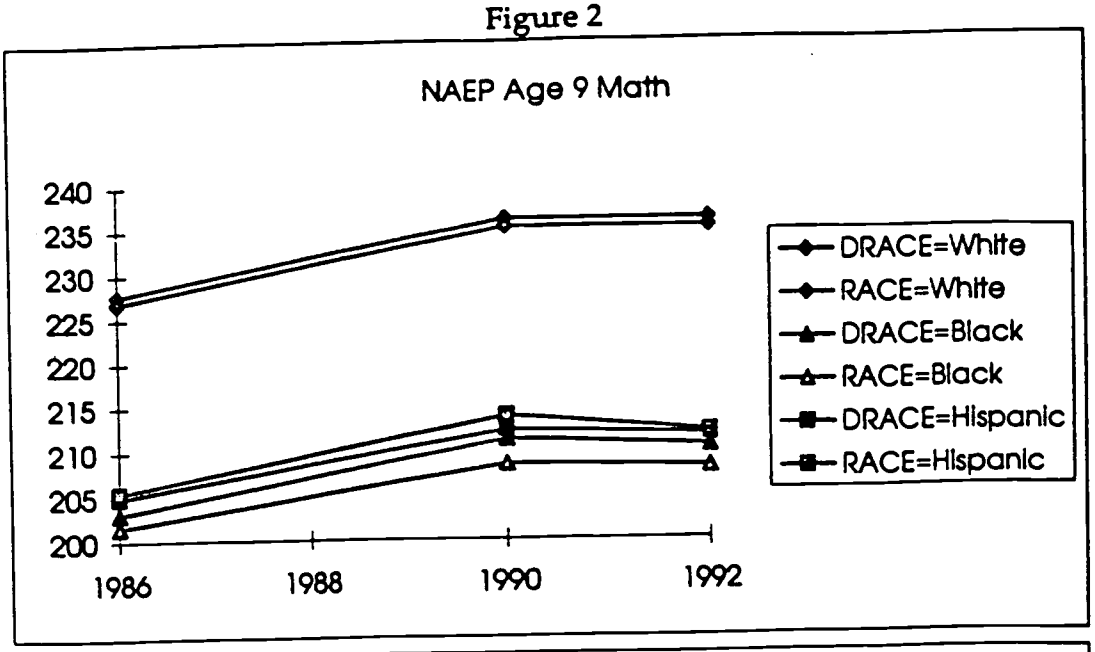
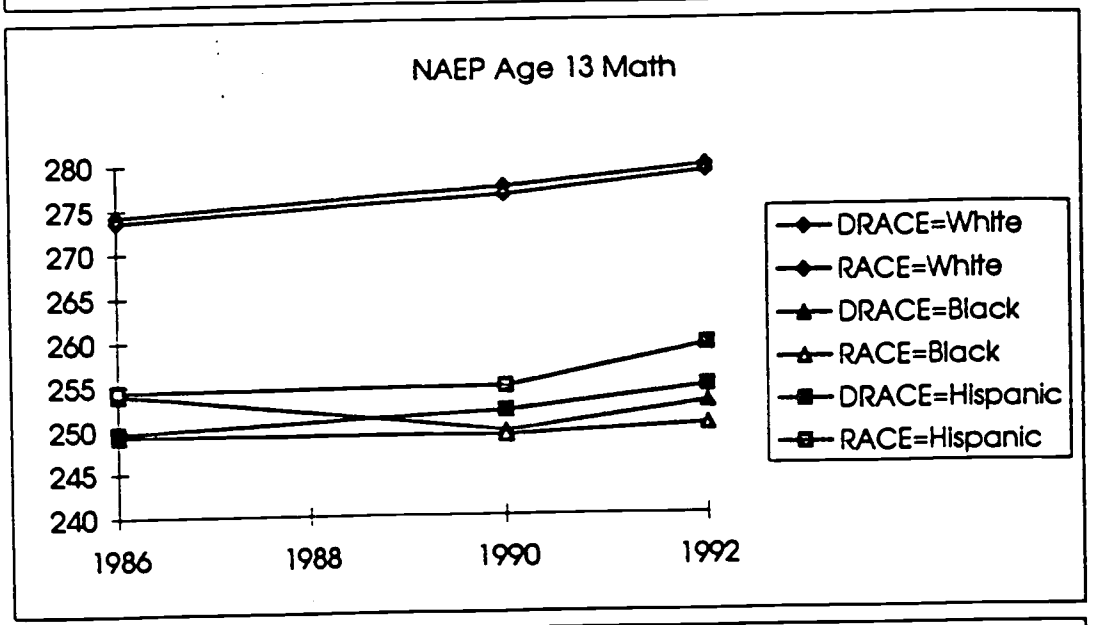


Figure 2


NAEP Age 9 Math
240
235
230
225
220
215
210
205
200
1986
1988
1990
1992
DRACE=White
RACE=White
DRACE=Black
RACE=Black
DRACE=Hispanic
RACE=Hispanic


NAEP Age 13 Math
280
275
270
265
260
255
250
245
240
1986
1988
1990
1992
DRACE=White
RACE=White
DRACE=Black
RACE=Black
DRACE=Hispanic
RACE=Hispanic

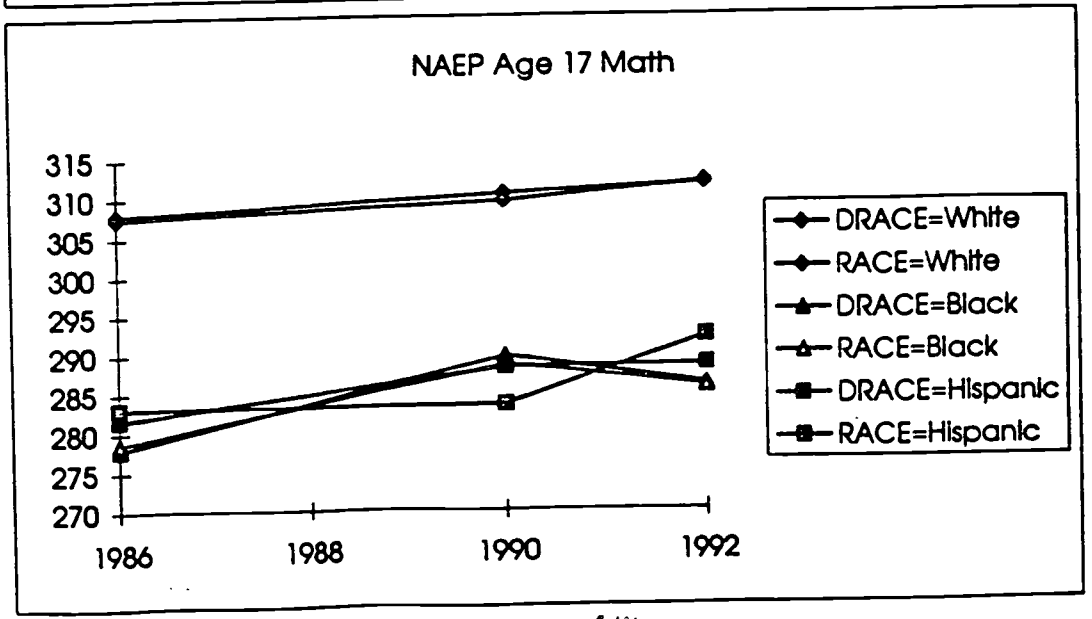

NAEP Age 17 Math
315
310
305
300
295
290
285
280
275
270
1986
1988
1990
1992
DRACE=White
RACE=White
DRACE=Black
RACE=Black
DRACE=Hispanic
RACE=Hispanic

Figure 2 (cont.)

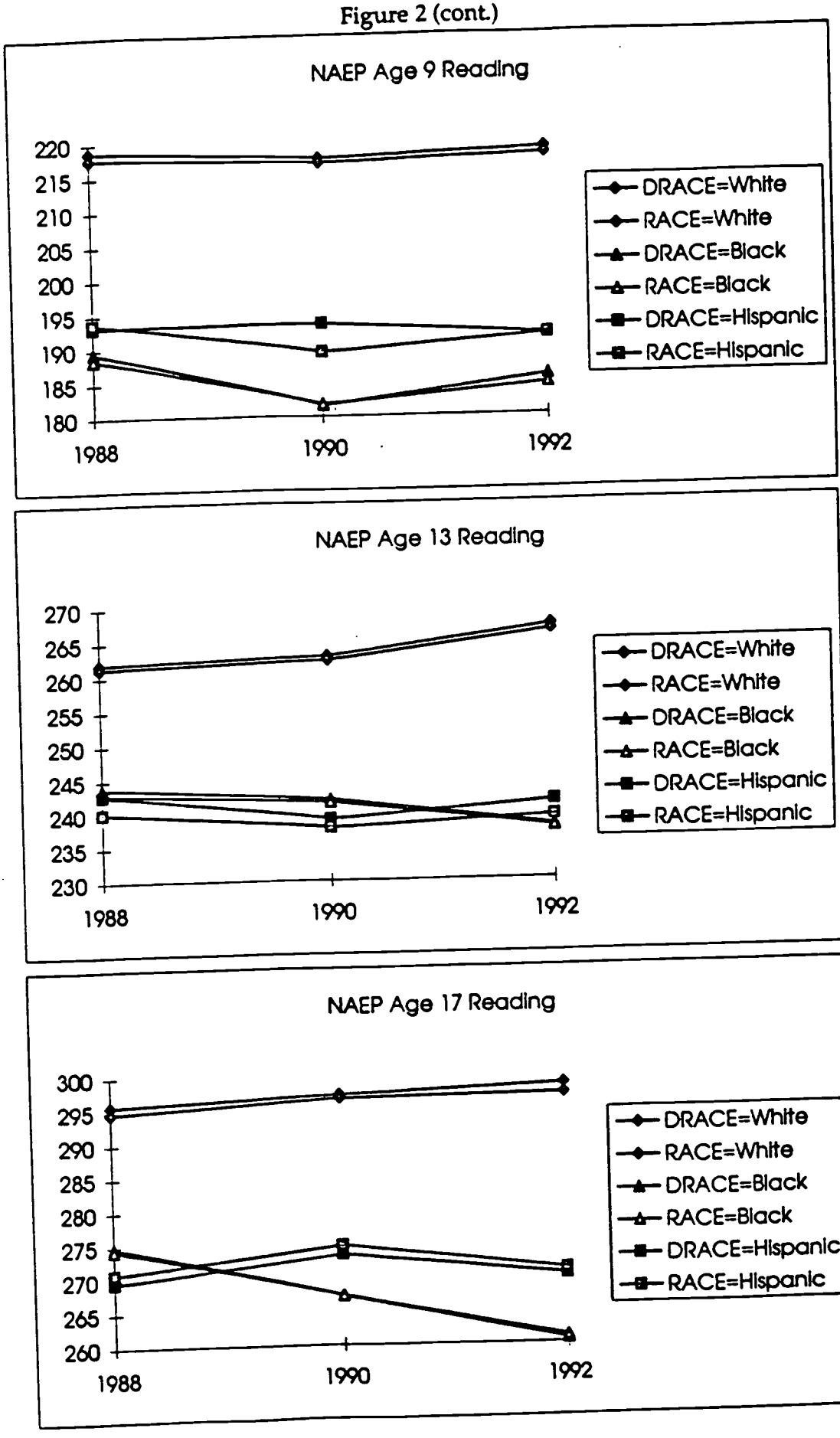


Figure 3

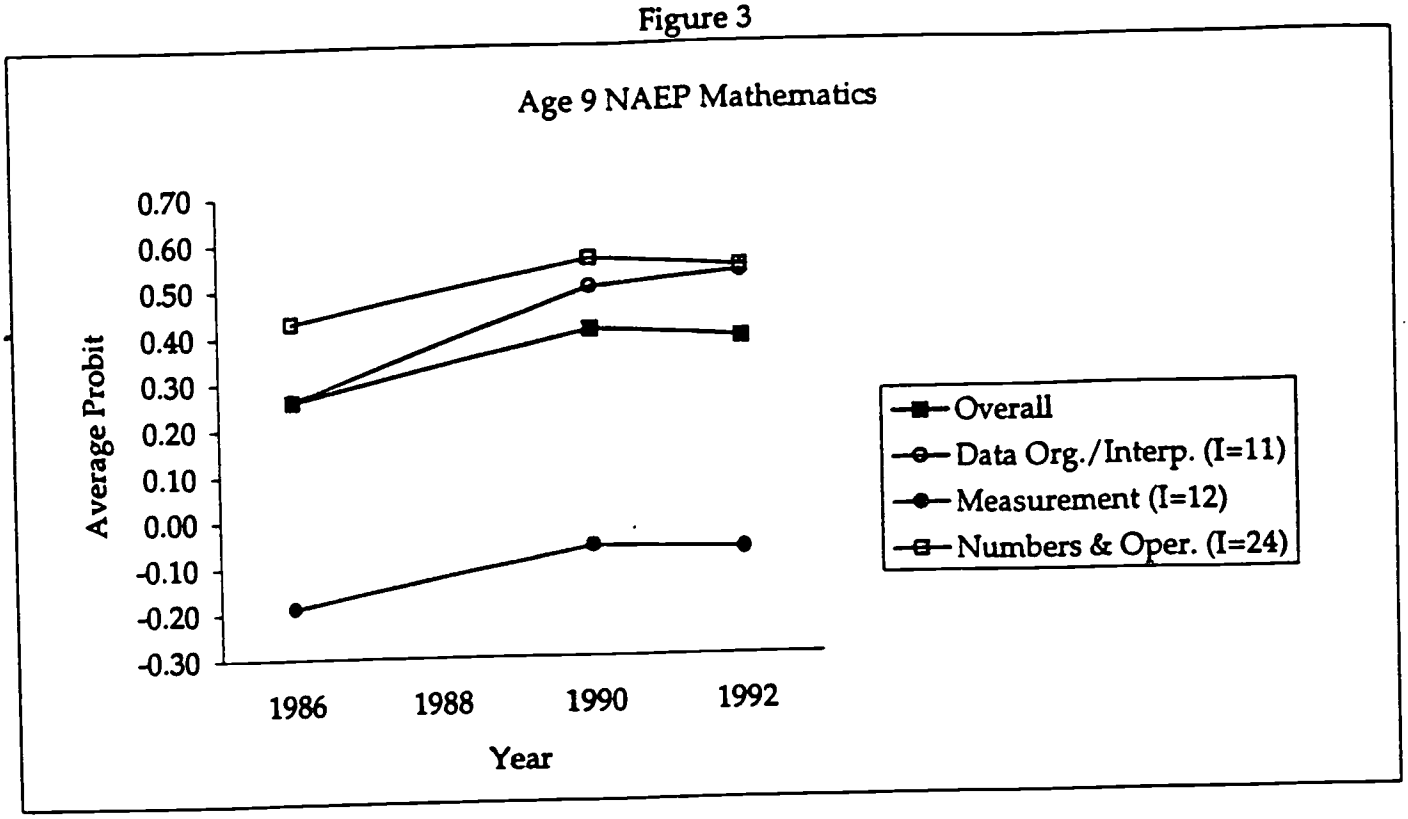

Age 9 NAEP Mathematics
Average Probit
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-0.10
-0.20
-0.30
1986
1988
1990
1992
Year
Overall
Data Org./Interp. (I=11)
Measurement (I=12)
Numbers & Oper. (I=24)

Figure 3 (cont.)

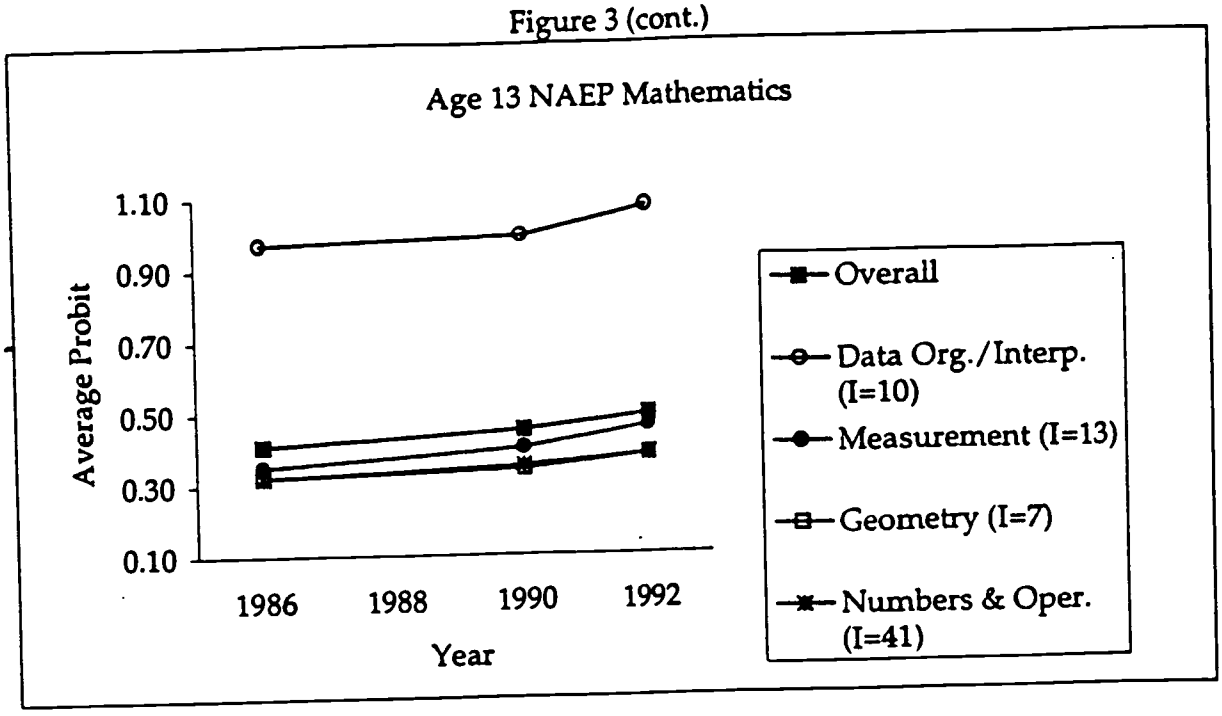


Figure 3 (cont.)

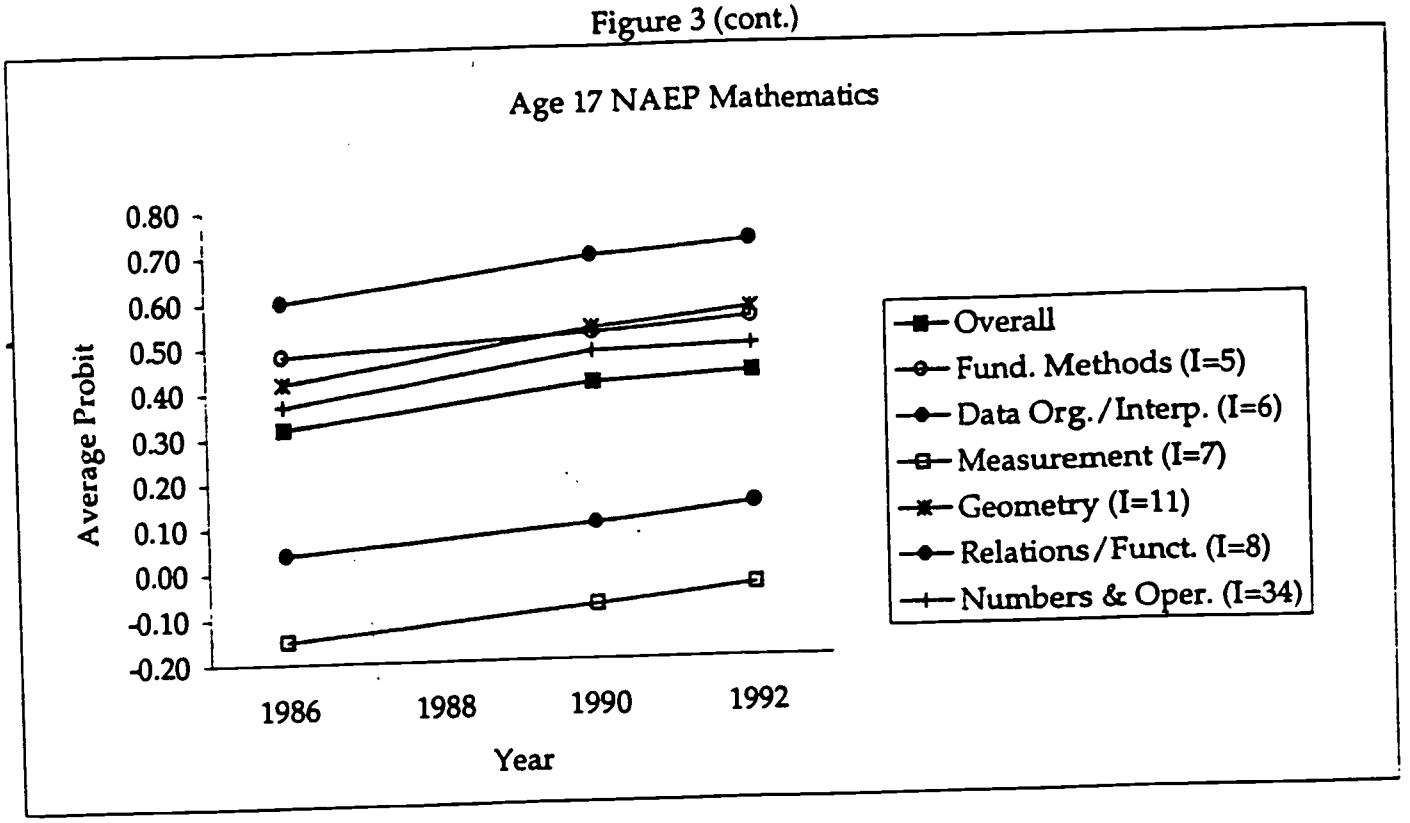


Figure 4

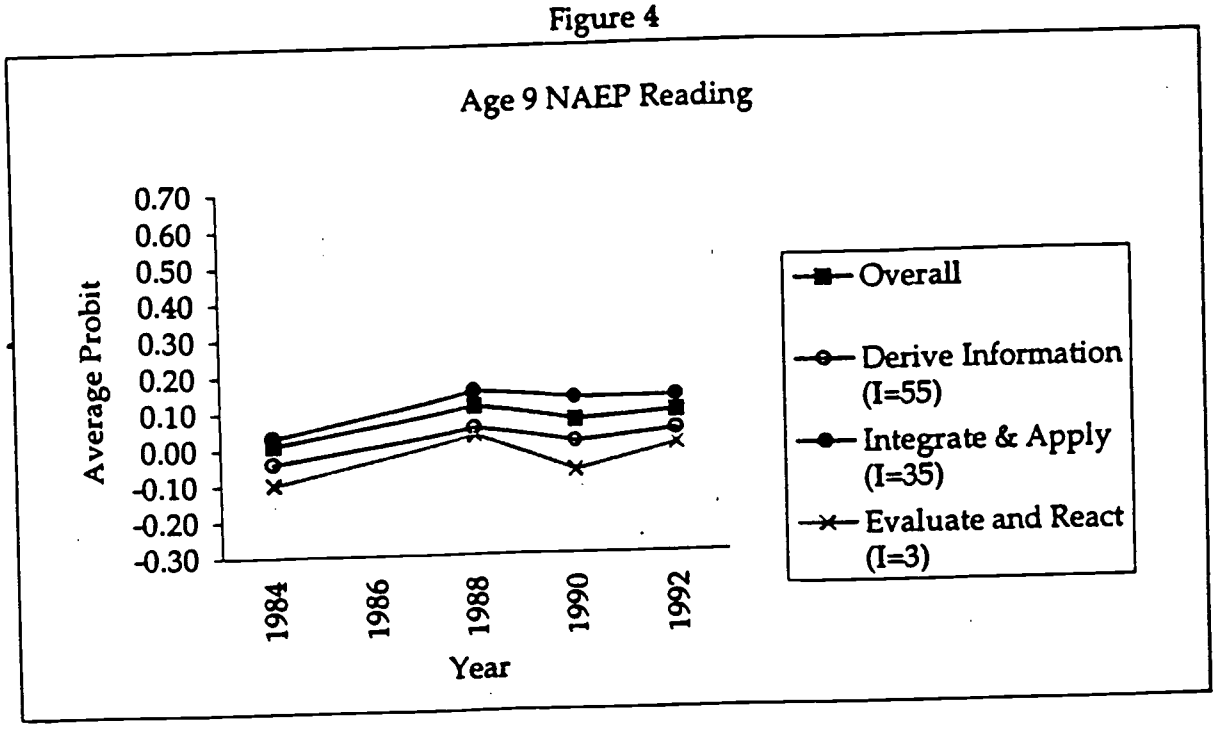

Age 9 NAEP Reading
Average Probit
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-0.10
-0.20
-0.30
1984
1986
1988
1990
1992
Year
Overall
Derive Information (I=55)
Integrate & Apply (I=35)
Evaluate and React (I=3)

Figure 4 (cont.)

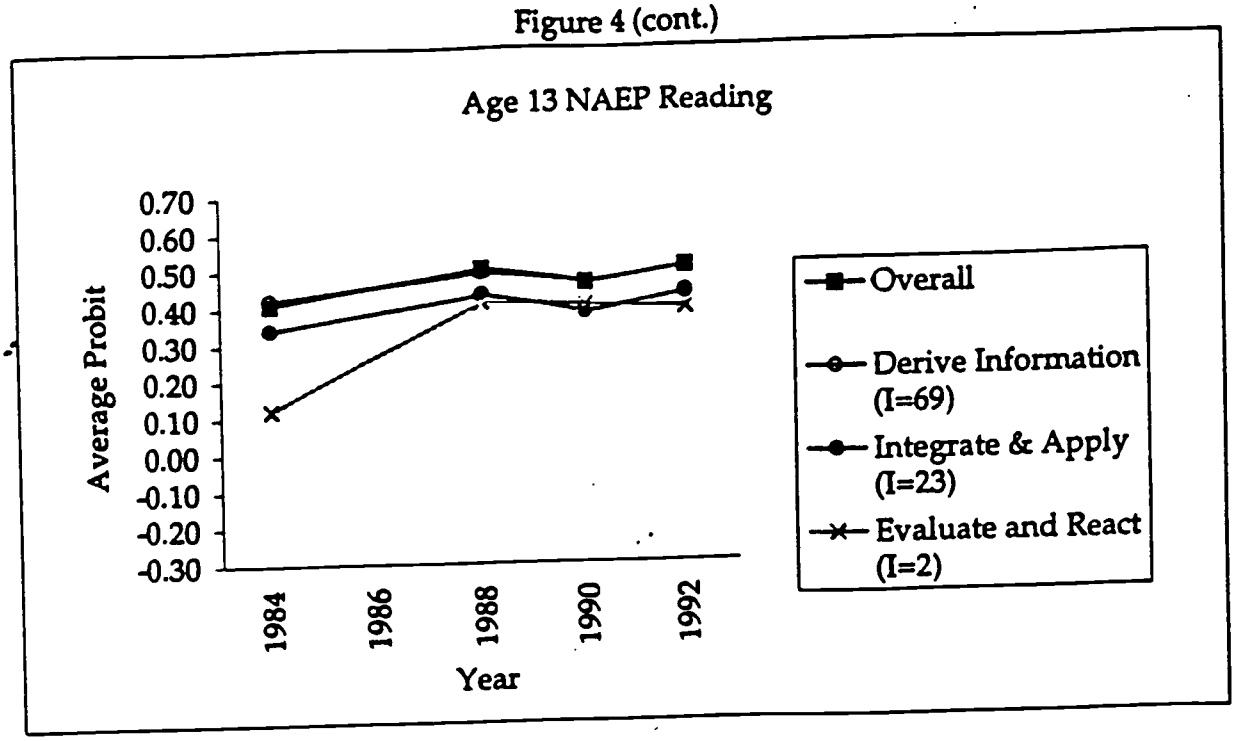


Figure 4 (cont.)

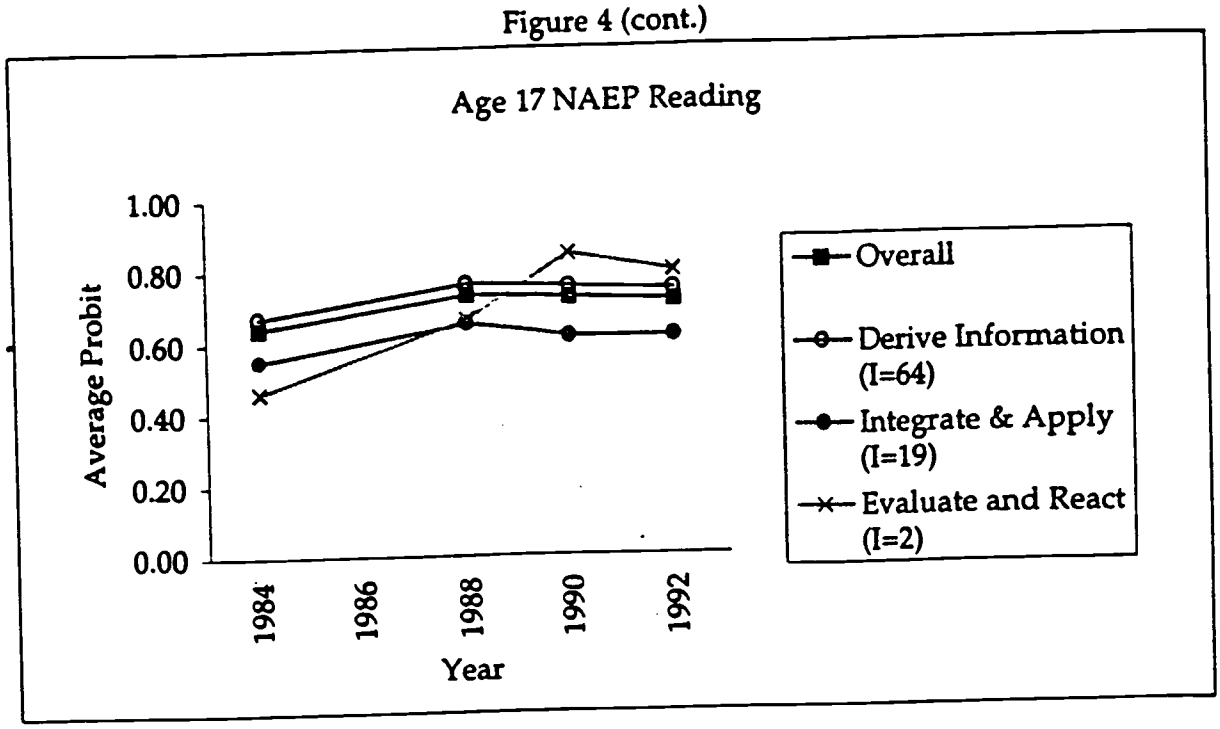


Figure 5

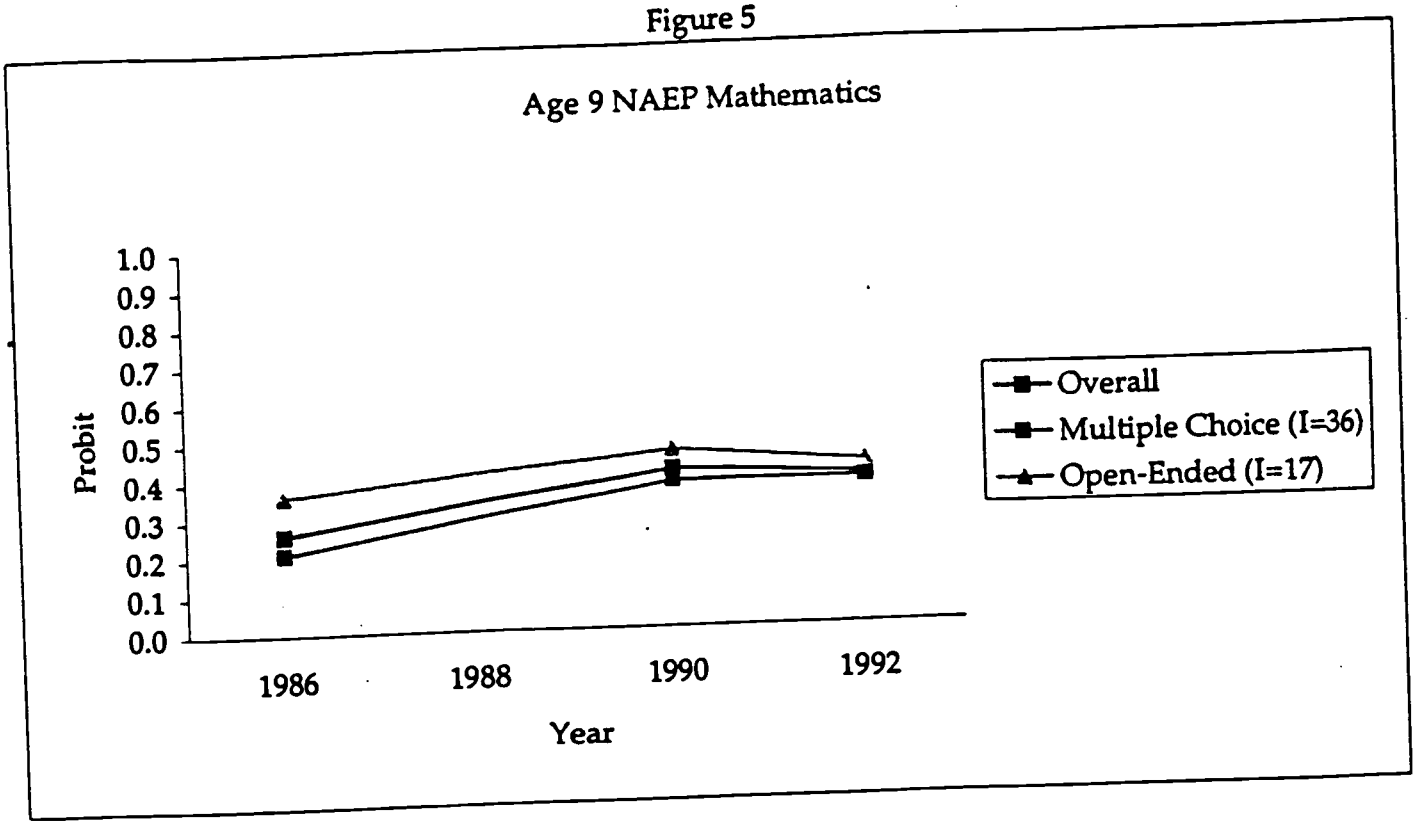

Age 9 NAEP Mathematics
Probit
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1986
1988
1990
1992
Year
Overall
Multiple Choice (I=36)
Open-Ended (I=17)

Figure 5 (cont.)

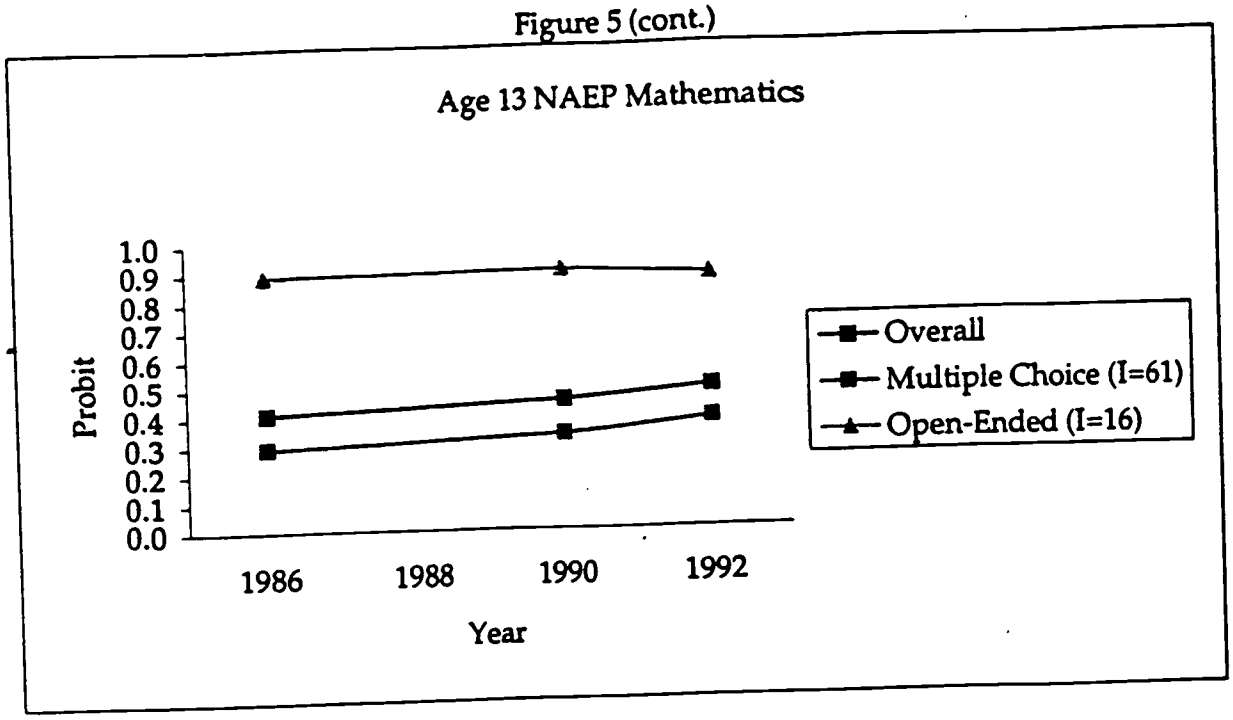


Figure 5 (cont.)

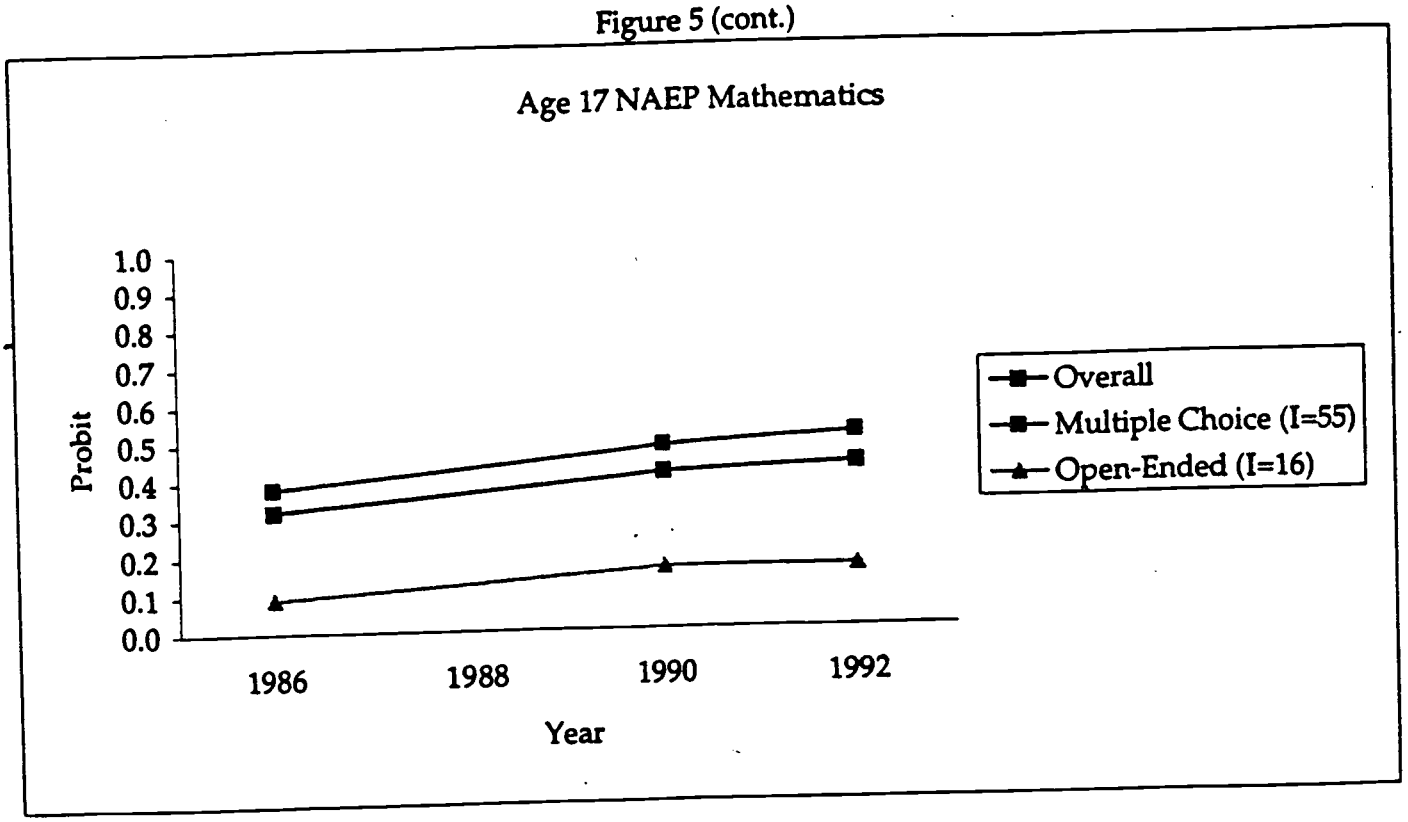


Figure 6

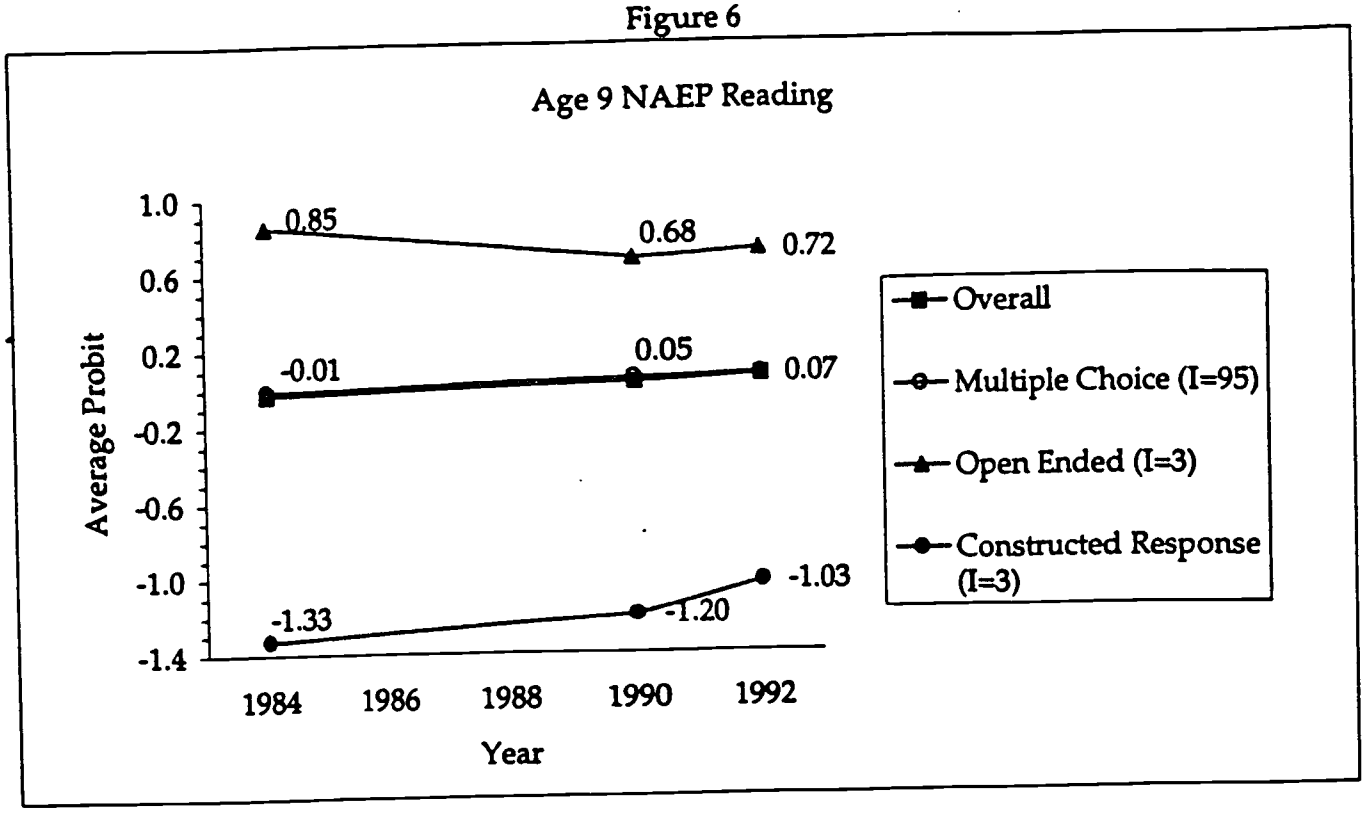

Age 9 NAEP Reading
Average Probit
1.0
0.6
0.2
-0.2
-0.6
-1.0
-1.4
0.85
0.68
0.72
-0.01
0.05
0.07
-1.33
-1.20
-1.03
1984
1986
1988
1990
1992
Year
Overall
Multiple Choice (I=95)
Open Ended (I=3)
Constructed Response (I=3)

Figure 6 (cont.)

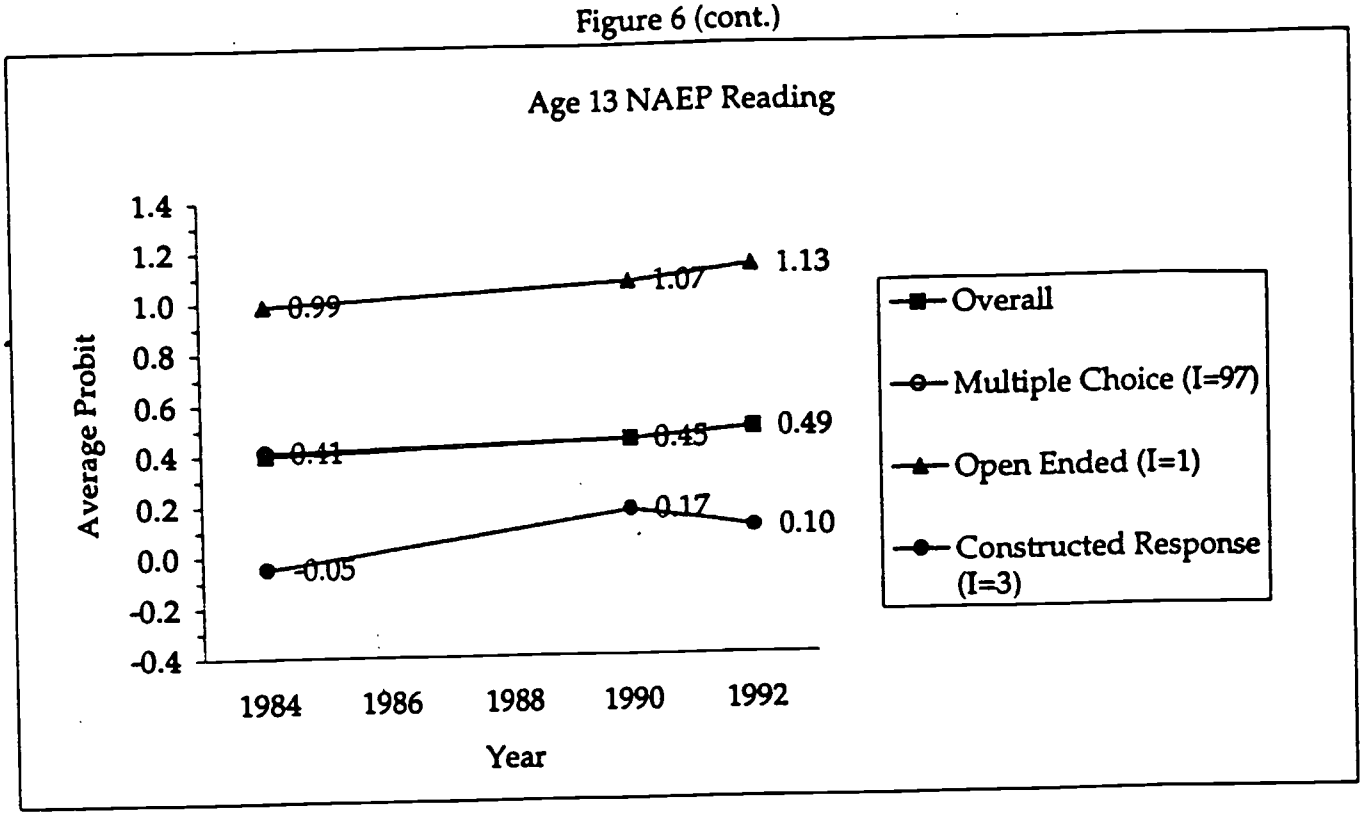


Figure 6 (cont.)

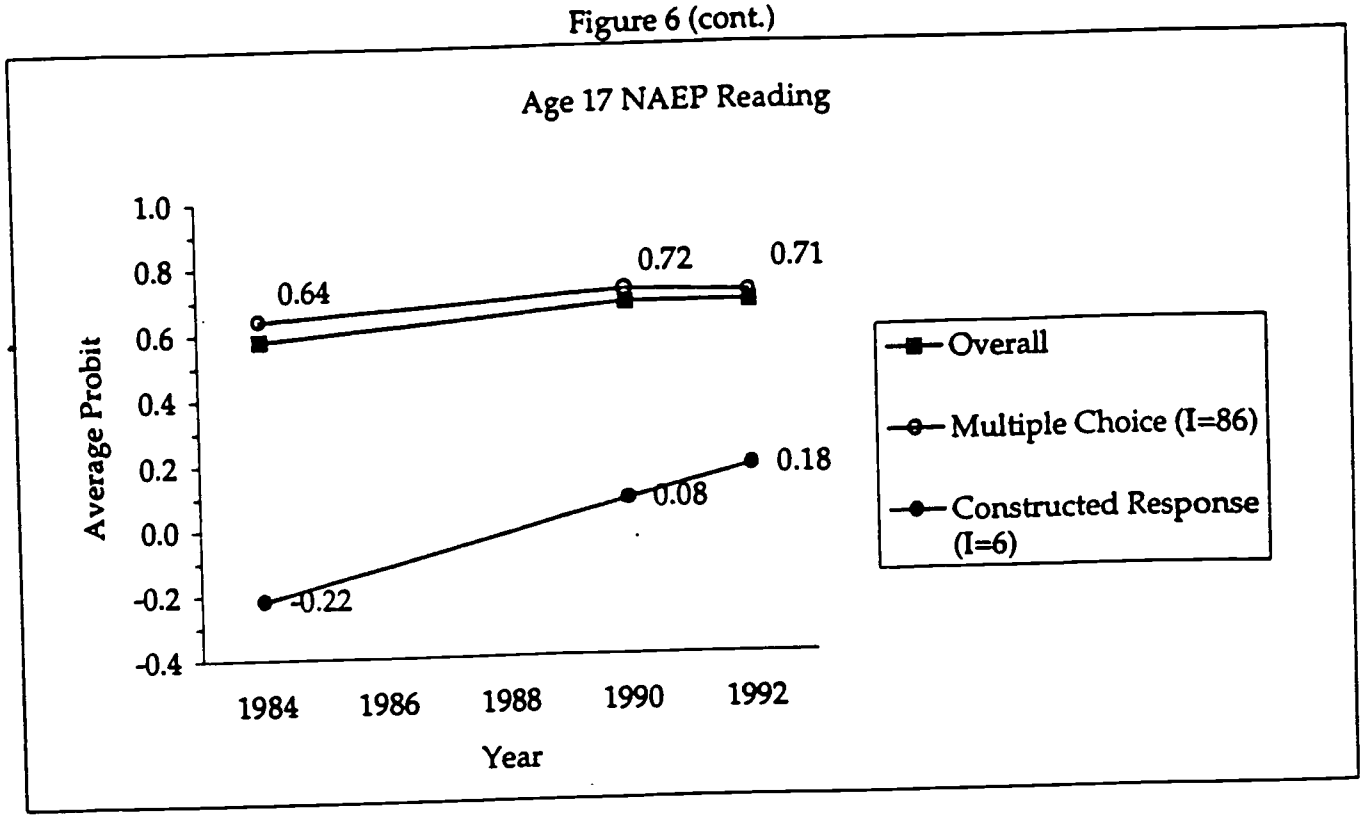


Figure 7

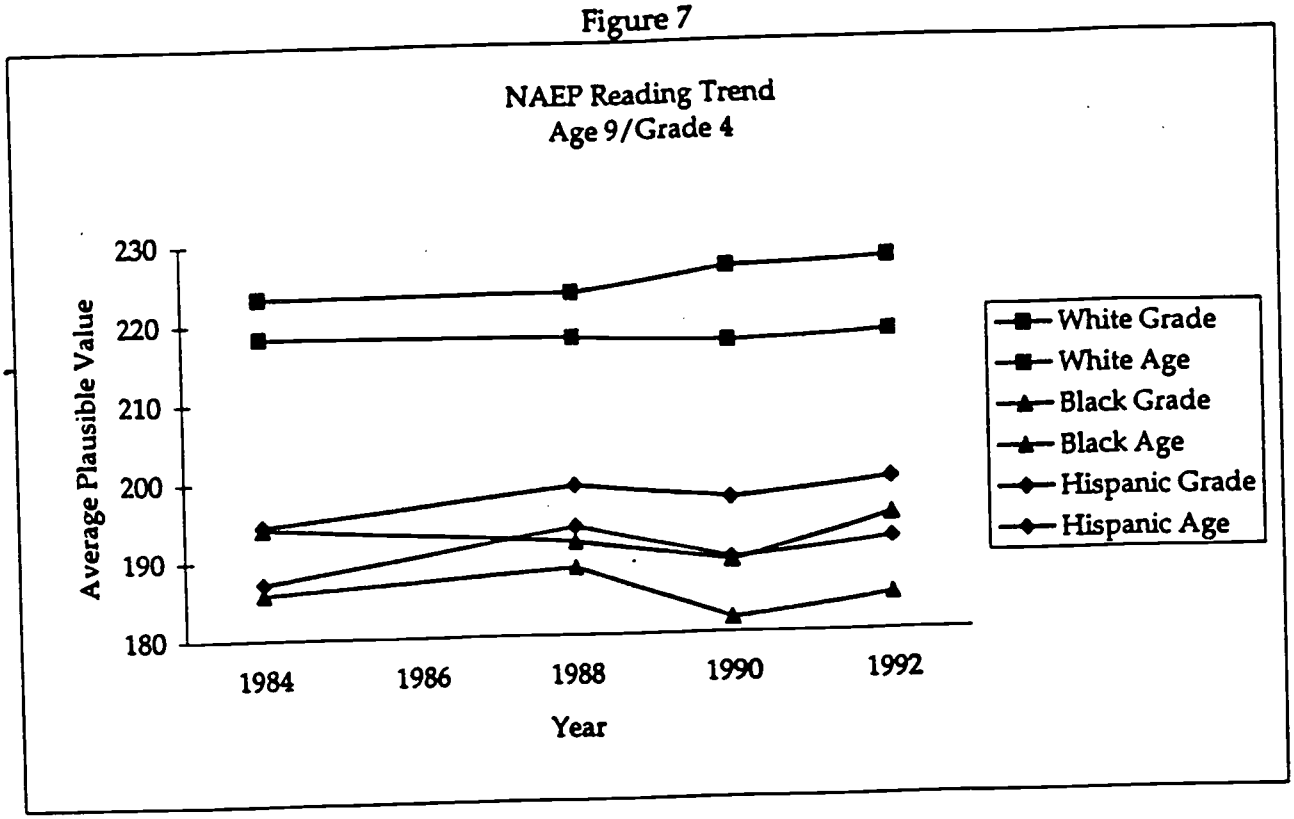

NAEP Reading Trend
Age 9/Grade 4
Average Plausible Value
230
220
210
200
190
180
1984
1986
1988
1990
1992
Year
White Grade
White Age
Black Grade
Black Age
Hispanic Grade
Hispanic Age

Figure 7 (cont.)

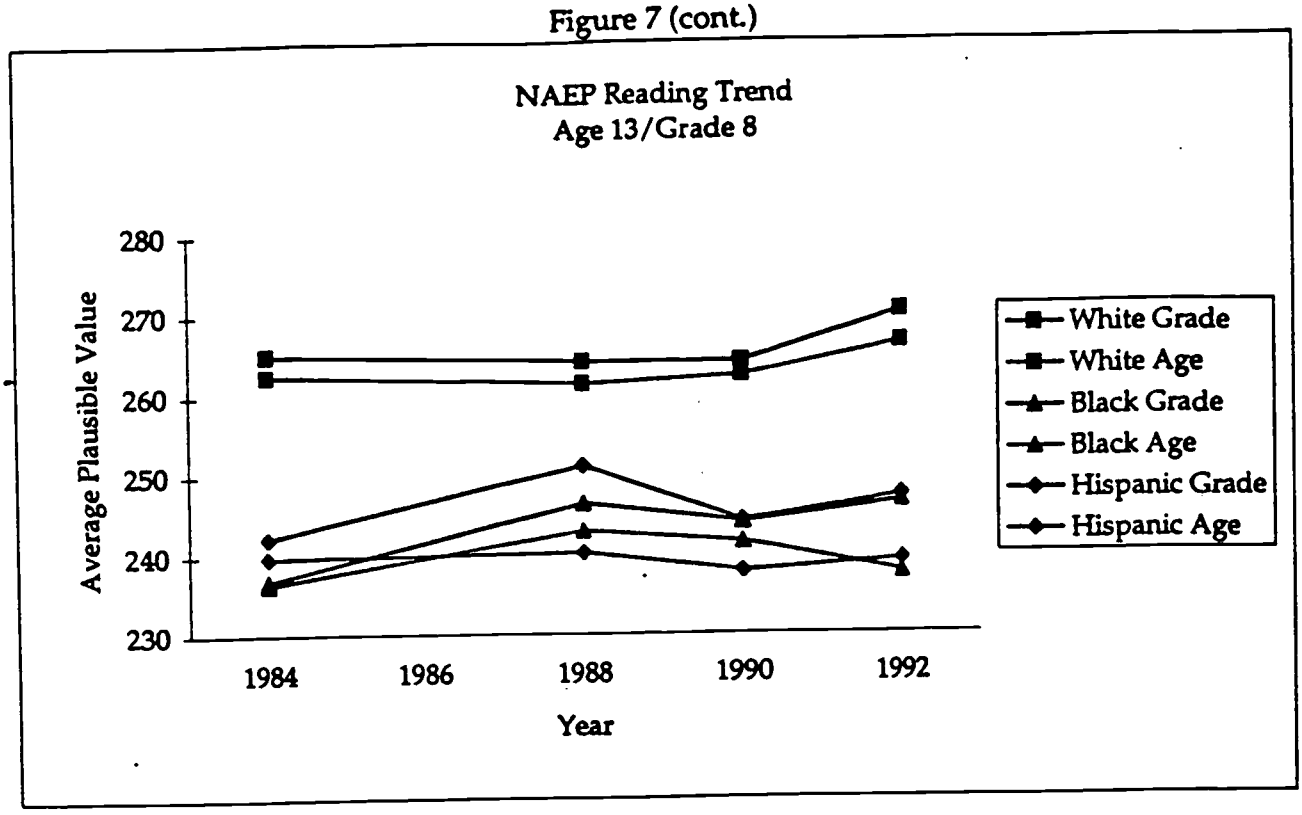
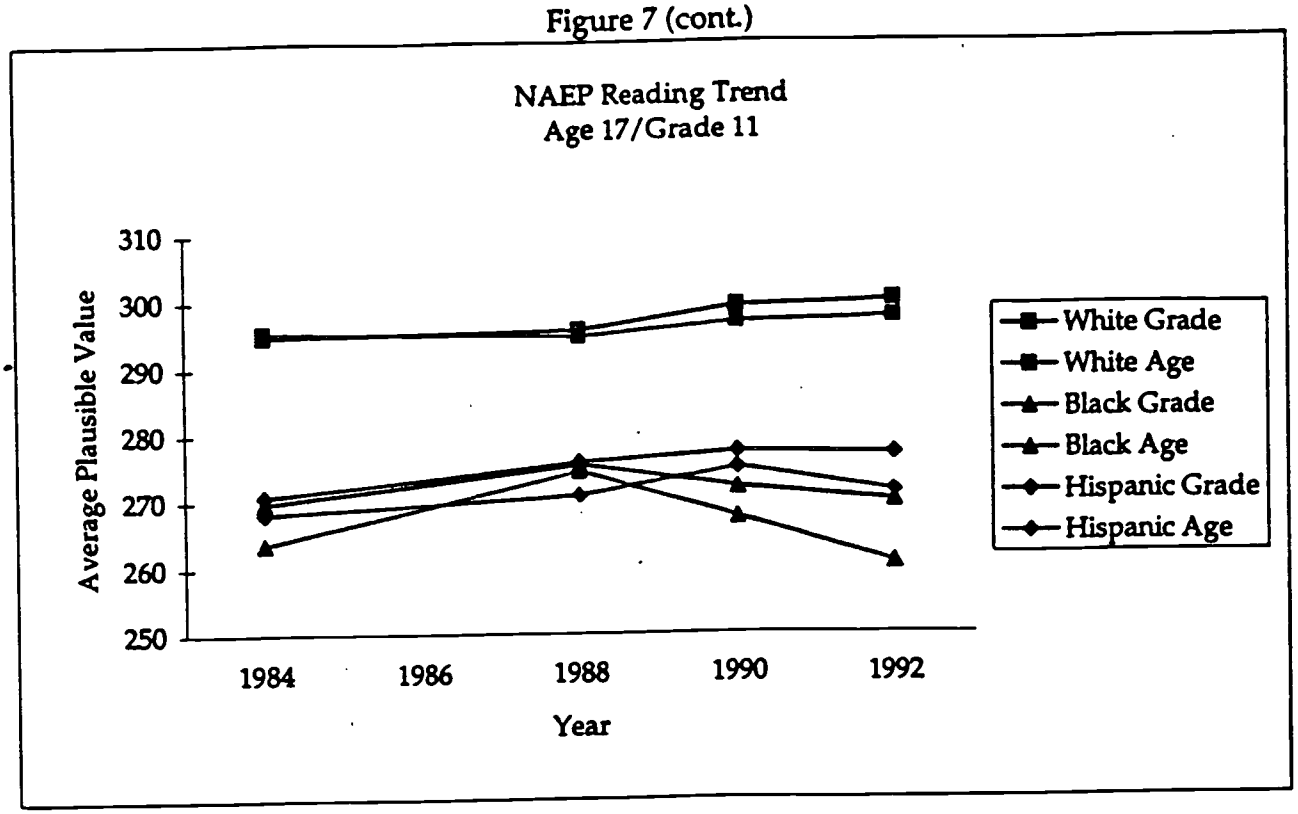
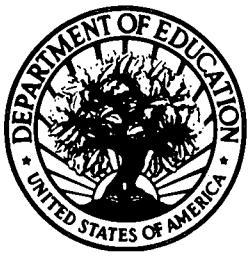


Figure 7 (cont.)





U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

NOTICE

REPRODUCTION BASIS

☐ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☒ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").